COLLOID CHEMISTRY

THEORETICAL AND APPLIED

Volume VII

LIST OF CONTRIBUTORS

A. E. ALEXANDER

JEROME ALEXANDER

GEORGE ANTONOFF

FREDERICK BERNHEIM

RICHARD J. BLOCK

DAVID M. BONNER

H. C. BORGHETTY

A. M. BUECHE

P. H. CARDWELL

A. CHWALA

THOMAS HILL DAUGHERTY

P. DEBYE

ARTHUR K. DOOLITTLE

EARL W. FLOSDORF

L. S. FOSDICK

JOHN W. GOWEN

LÉON GRIGORAKI

WILLIAM D. HARKINS

ERNST A. HAUSER

J. C. HENNIKER

L. J. E. HOFER

IRVING M. KLOTZ

ALFONS KRAUSE

PAUL G. LAUFFER

D. S. LE BEAU

H. W. LEVERENZ

DUNCAN MACRAE

W. J. MATTOX

J. W. MCBAIN

F. J. NELLENSTEYN

C. R. OSWIN

AMAR NATH PURI

D. S. P. ROEBUCK

CHARLES M. ROSSER

HERBERT A. SOBER

CHARLES A. SOUTHWICK, JR.

ALFRED J. STAMM

C. A. STOKES

DONALD STOOKEY

HAROLD TARKOW

R. N. TRAXLER

H. W. VAHLTEICH

E. J. W. VERWEY

ARTTURI I. VIRTANEN

ALEXANDER S. WIENER

COLLOID CHEMISTRY

THEORETICAL AND APPLIED

By Selected International Contributors

COLLECTED AND EDITED BY

Jerome Alexander

Volume VII

THEORY AND METHODS
BIOLOGY AND MEDICINE
TECHNOLOGICAL APPLICATIONS

PREFACE

T<small>HIS VOLUME INCLUDES</small> papers on theory and methods, biology and medicine, and technology. Many of the papers indicate increasing realization of the fact stressed throughout this series that there are quite a number of successive levels of material structure, and that the properties of any unit emerging at any level, while dependent upon the nature and arrangement of its subsidiary constituent units, may nevertheless exhibit entirely new emergent behaviors. Experimental data of all kinds should be considered in this light. The papers in the several volumes have been so chosen and arranged that readers are exposed to the results of research in widely different fields and sometimes even to apparently conflicting results in the same field. Many mysteries and seemingly irreconcilable differences vanish when we acquire depth as well as breadth of mental focus. For example, Pasteur was correct in attributing fermentation to microorganisms; but Liebig and Bertholet were also correct, as Buchner demonstrated in 1895, in attributing fermentation to what we now know to be the much smaller enzyme-catalysts formed within the microorganisms.

The ever-recurring development of characteristic individuals from seeds and eggs led Aristotle to apply the word "entelechy" (i.e., having within a distant objective) to describe what he saw. This term has long been *metaphysical* to scientists, for it neither gave nor suggested any explanatory mechanism. But as scientific understanding reached into lower and lower levels of structure, it became evident that the basic chemical changes in living things are dominated by biocatalysts which constitute a *material catalyst entelechy*. Operating in a suitable milieu and framework, biocatalysts normally determine that the right substances are formed in the right proportions, at the right places, and at the right times, so that, barring abnormalities, definite chemical outputs and consequent structures are forthcoming. Though the basic chemical changes occur at invisible atomic and molecular levels, the end results are usually read at visible levels by embryologists, zoologists, botanists, and clinicians,

who must use instruments to peer down into lower levels in order to understand the mechanisms producing the results they see, and to discover if and how these results may be controlled or modified. The term "entelechy" is thus transported from the limbo of metaphysics to the free upper air of experimental science. Similarly, much of our technological progress stems from our understanding of the chemistry of practical processes and the utilization of catalysts to facilitate known chemical changes or to produce new ones.

Danger to science does not come from those who advance new, sometimes startling, and often erroneous ideas which are subjected to the free and open criticism of fellow scientists throughout the world. False notions are usually quickly scotched, though there is sometimes an understandable but unfortunate tendency for the graduates of institutions to maintain, teach, and thus perpetuate not only the truths but also the errors of their superiors, and in anticipation of future advancement, to "crook the pregnant hinges of the knee, where thrift may follow fawning." Sound innovations, if they do not come from scientifically orthodox sources, and especially if they militate against the views currently maintained and taught by scientists in high positions and publicized by press agents or by science writers, often die a-borning. They are sometimes denied publication, or if published, may be ignored in the writings and bibliographies of the elite. Later on they may be claimed as discoveries by others, even by referees who, as a reviewer once keenly put it, have the unfortunate habit of remembering what they have read but of forgetting that they have read it. But the truth finally burns through.

The greatest danger to science comes from those who aim to subordinate accurate observations and sound reasoning to preconceived scientific, theological, sociological, or political ideas. Roger Bacon and Galileo were imprisoned by the church, and Servetus was burned at the stake by Calvin, for supporting unorthodox scientific ideas. The ludicrous Scopes case in Tennessee (U.S.A.), where a state banned the teaching of evolutionary theory as contrary to theological dogma, lights up the fact that in most civilized states this type of persecution of science is on the wane; but in Russia the teaching of modern genetics has been officially banned. On the other hand, geneticists intent on proving the basically important functions of genes and chromosomes, looked askance at cytoplasmic determinants, now recognized as important in disease and in normal and abnormal development. The establishment of large, endowed, corporate and governmental research institutions places great powers within the hands of a few executive heads who may show any of the usual human traits, admirable and otherwise. Where ignorant, selfish, prejudiced, or venal personalities dominate, scientific advance may suffer. The ideals of science often transcend the practices of some scientists.

The experimental work of science consists in observing what happens when the positions of some material units change, or are changed. From the observations thus made, we then make deductions and draw conclusions. Science, therefore, is based on the existence and inherent powers of matter, as well as on the existence and proper operation of rational human minds.

The evidence of our senses is that from material units there have evolved a multitude of structures, among them a great variety of self-duplicating living things, some of which have also evolved sensitivity, mentality, and finally, reasoning power. But the ultimates of matter and the origin of the wonderful universe we study remain a mystery. No one has ever actually created any material unit. We often speak of the creations of chemists, physicists, architects, and builders, and even of the "creations" of dressmakers; but materially these involve simply a rearrangement of existing units of some kind, even though the patterns and results are new to our mentality. We have no experimental approach to the creation of any unit temporarily considered to be "fundamental." Even conceptually, creation is inscrutable.

At present, our fundamental units of matter are electrons, positrons, protons, and neutrons; and the "spin" of some of these bespeaks their complexity. Mesons (but yesterday called mesotrons) and polyelectrons have a transient existence. While the photon, gravitron, and neutrino are spoken of as "units of energy," and we deal mathematically with electric, magnetic, and gravitational "fields," many are content to shrug off the physical meaning of these terms as "metaphysical" or "philosophical," little realizing that what we call science is grounded on metaphysical and philosophical assumptions. But the frontiers of science are ever extended by those who are not content with dogmas and textbooks worshipped and sometimes promulgated by Bourbons of science who "learn nothing and forget nothing." Speaking of the exploration of the atom by Lord Rutherford, Stephen Miall wrote:

> *He made plain the invisible;*
> *He broke up the indivisible;*
> *He changed the immutable;*
> *And he unscrewed the inscrutable.*

In what turned out to be his last address, Lord Rutherford said, with becoming modesty: ". . . it is not within the nature of things for any one man to make a sudden violent discovery; science goes step by step, and every man depends on the work of his predecessors. When you hear of a sudden unexpected discovery—a bolt from the blue, as it were—you can always be sure that it has grown up by the influence of one man on

another, and it is this influence which makes the enormous possibility of scientific advance. Scientists are not dependent upon the ideas of a single man, but on the combined wisdom of thousands of men, all thinking of the same problem, and each doing his little bit to add to the great structure of knowledge which is gradually being erected."

JEROME ALEXANDER

New York, New York
January, 1950

CONTENTS

Part I

THEORY AND METHODS

THE INTERMEDIATE-COMPOUND NUCLEUS
IN NUCLEAR REACTIONS *

William D. Harkins

University of Chicago, Chicago, Ill.

Fundamental Concepts Regarding the Nucleus

Synthesis of Nuclei. In 1926, the writer introduced into nuclear science an entirely new concept: that what had been considered as an artificial *disintegration* is actually the artificial *synthesis* of a new, heavier nucleus. This nucleus disintegrates after a certain life period, which is usually short.

In 1919, Rutherford found that the long-range particles, earlier shown by Marsden to be produced when α-particles are shot into hydrogen, are hydrogen nuclei (now designated as protons) and showed that the range is very much longer when nitrogen instead of hydrogen is bombarded. This indicated that a nuclear disintegration had occurred in which one product was the proton. The natural assumption was that the kinetic energy of the fast α-particle was so large that the nitrogen nucleus was broken into two parts by the collision, or

$$_7N^{14} \longrightarrow {_6}C^{13} + {_1}H^1$$
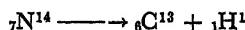
Later this became known as disintegration by noncapture. The writer considered that the mechanism might not be as simple as this and resorted to photography to determine what the process actually is. The results of this work indicated to the writer that the *primary* process is a nuclear synthesis.

Composition of the Nucleus. Any atomic nucleus consists of P positively-charged particles, or protons, and N uncharged particles, or neutrons. Both particles are designated as nucleons. The mass of the proton is 1.008123 and of the neutron 1.00893, when the nucleons are free. However, the mean mass of nucleons becomes close to 1.000 when bound

3

into nuclei heavier than that of nitrogen ($m = 14.00751$), as compared with the mass of the most abundant isotope of oxygen taken as 16.000000.

From the standpoint of ordinary chemistry an important feature of the nucleus is the number (P) of protons (p^+) which it contains, since by the principle of electrical neutrality the number of electrons in the outer atom must be equal to the number of protons in the nucleus. The chemical and physical properties of material are related to the number of electrons in the atom, and exhibit a periodic relation expressed in the decade 1860-70 by the periodic system of Mendelejeff and others. The principal periods of this system are expressed by the numbers 2, 8, 18 and 32.

An entirely different periodic system of atomic nuclei was developed by the writer between 1915 and 1923. Here the principal period is 2, as in the outer atom, but the other *special numbers* related to nuclear properties are different from those outside the nucleus. This system will be considered below.

Packing Effect and Heat of the Sun and Stars. In 1915 and 1921, Harkins and Wilson advanced the theory that the source of heat in the sun and stars is a nuclear reaction in which four atoms of hydrogen are transformed into one of helium. The mass of the hydrogen atom as given above is 1.008123, or that of four hydrogen atoms, 4.032492. The mass of the helium atom is 4.00390, so that the loss of mass in the transformation of hydrogen into helium is 0.03159 gram for 1 gram-atom of helium. According to the special relativity theory of Einstein this loss may be transformed into units of ergs of energy by multiplying by c^2, where c is the velocity of light (2.99776×10^{10} cm per sec). Thus

$$E = \Delta mc^2 = 0.03159 \times (2.99776)^2 \times 10^{20}$$
$$= 0.28388 \times 10^{20} \text{ ergs}$$
$$= 6.7816 \times 10^{11} \text{ mean calories}$$

is the heat liberated by the formation of four grams of helium. It was thereby shown that the heat developed in the sun and stars is, per pound of hydrogen consumed, equal to that produced by burning 10,000 tons of coal. In the atomic bomb, a pound of uranium 235 gives as much energy as the combustion of 1500 tons of coal. Thus hydrogen, per unit weight, gives about seven times more energy than uranium 235.

Either a proton or a neutron in a nucleus is designated as a nucleon. Since the average mass of a nucleon in a nucleus is very close to 1.000, what is known as the integral mass A gives the number of nucleons in the nucleus. The largest number of nucleons in any stable nucleus is 209 in bismuth, but this rises to 242 in curium, which is unstable and gives off α-particles. The smallest number of nucleons in any stable

nucleus is *one* for the proton. It is also *one* for the neutron, which is not stable as a separate entity.

Size of the Nucleus. The radius of the proton or of a neutron is about 1.5×10^{-13} cm. Since in a complex nucleus the nucleons are closely packed, the approximate radius is given by

$$R = 1.5 \times 10^{-13} A^{1/3} \text{ cm}$$

If calculated on this basis, the radius of the bismuth nucleus is found to be 7.9×10^{-13} cm, and the diameter, 1.58×10^{-12} cm. The unit 10^{-12} cm, when squared, gives 10^{-24} cm ^{2} per nucleus, which is now designated as a *born*, so that:

$$1 \text{ born} = 10^{-24} \text{ cm}^2/\text{nucleus}$$

and this unit is used in designating the cross section of a nucleus. A very common diameter for a moderately small molecule is 3×10^{-8} cm, or the volume is of the order of 27×10^{-24} cm^3.

If we take the volume of the oxygen nucleus as 1.7×10^{-37} cm^3, the volume given above for an atom is about 10^{14} larger. Thus the density of nuclear material is of the general order of 10^{14} times larger than that of ordinary solid substances. The density of a nucleus is of the order of 5×10^{14}, or one cu cm of neutrons and protons, *if these could be tightly packed* as spheres, would have a mass of 3.7×10^{14} grams of 370 million tons (metric) or 3.7×10^{14} tons per cu meter. Thus a cube with an edge of 250 meters would have the mass of the earth, or a lady's thimble would hold about 200 million tons of nuclear material. The thimble, on this basis, would hold the material of 2300 great ships of the mass of the Queen Elizabeth, the largest vessel in the world.

Intermediate-Compound Nucleus Formation and Its Consequences

Synthesis and Disintegration as Separate Processes. The early work (1919-23) of Rutherford on the artificial disintegration of the nucleus was carried out by observing the scintillations on a zinc sulfide screen caused by long-range protons. This gave no knowledge of the mechanism of the process and the natural assumption was made that the kinetic energy of the fast α-particles was sufficient to split the nucleus into two fragments. In 1923, Harkins and Ryan [1] suggested a method by means of which the mechanism could be determined. This consisted in bombarding nitrogen, or another gaseous material, in a Wilson cloud-track apparatus, with very fast α-particles from thorium C′ with a velocity of $0.0688\ c = 2.06 \times 10^9$ cm sec^{-1}. Table 1 indicates that the highest kinetic energy for any of the α-particles listed is less than $\frac{1}{3}$ the energy of formation of an α-particle from protons and electrons. The velocity

of light is expressed by c. Thus the α-particle from the thorium C′ used in the work at the University of Chicago has an initial velocity of 12,200 miles per second.

TABLE 1. ENERGY OF ALPHA-PARTICLES

Energy of formation of 1 gram-atom of helium from 4 gram-atoms of hydrogen
= 2.8×10^{19} ergs (-4.6×10^{-5} ergs per atom of He)

Source of α-Particle	Velocity (c)	Kinetic Energy	
		(per particle $\times 10^5$)	(per gram-atom $\times 10^{-18}$)
Po	0.0523	0.812	4.92
Ra C′	0.0641	1.218	7.38
Th C	0.0572	0.970	5.88
Th C′	0.0688	1.404	8.51

The first paper which considered that a nuclear disintegration is preceded by an artificial synthesis, written in 1926, antedated any other on this subject by ten years, during which time all theoretical physicists who considered the subject opposed the theory. Two steps were described in this paper: (1) The first step is the synthesis of a new and heavier nucleus, the whole-number mass of which is the sum of the whole-number masses of the two nuclei which unite, and the charge of which is equal to the sum of the charges of the two nuclei. (2) The disintegration of this nucleus is a later, independent event. The title of the paper was: "The Synthesis and Disintegration of Atoms as Revealed by the Photography of Wilson Cloud Tracks." Initially an attempt was made to find the track of the intermediate nucleus in the photographs, although it was recognized that this is in general impossible on account of its short life and resultant shortness of track.

The nucleus synthesized in the first step is designated as the *intermediate nucleus* because it disintegrates later and is an intermediate state between the initial and final nuclei. It is designated also as a *compound* nucleus to indicate that it is a compound formed by the collision and union of two lighter nuclei. The first synthesis investigated in this laboratory was that of F^{18} from N^{14} and He^4, or

$$_7N^{14} + {}_2He^4 \longrightarrow {}_9F^{18*} \tag{1}$$

where the asterisk indicates that the fluorine nucleus is in an excited state in which it contains an excess of energy. This will cause it to disintegrate later. The idea developed was that *this excited nucleus may now disintegrate in any one of a number of possible ways.* One possibility is:

$$_9F^{18*} \longrightarrow {}_8O^{17} + {}_1H^1 + \text{k.e.} \tag{2}$$
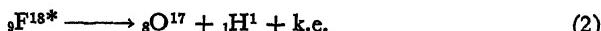
in which k.e. indicates the amount of energy liberated. Obviously it cannot disintegrate in any way which would increase the energy. This

type of disintegration is illustrated by Figures 1 to 4. Disintegration may also occur to give the initial particles:

$$_9F^{18*} \longrightarrow {}_7N^{14} + {}_2He^4 + k.e. \tag{3}$$
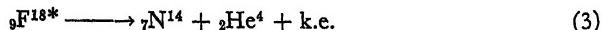
A third method of disintegration is:

$$_9F^{18*} \longrightarrow {}_8O^{17} + {}_0n^1 + e^+ + k.e. \tag{4}$$
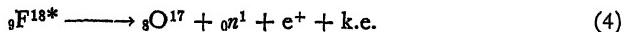
Figure 1. Two views at right angles of the nonelastic collision in which the intermediate nucleus $_9F^{18*}$ is formed. At the left of each view an α-particle impinges upon the nucleus of a nitrogen atom and unites with it. An extremely faint straight track extends horizontally to the right of the fork and is due to a proton or H-particle emitted at high speed. This has a range 19.6 cm in air at 15° and 760 mm pressure. The track of the oxygen atom which is formed slants upward to the left from the fork. The left-hand view is overexposed as a result of an irregularity in the lighting, but the track of the H-particle is easily visible on the original negative.

Characteristics of the Fork

Fork	ϕ	θ	$R\alpha$	$V\alpha + 10^{-9}$	$Vp + 10^{-9}$	Rp	$Vn \times 10^{-9}$	Rn	E_2/E_1
	118° 32'	15° 21'	6.3	1.86	2.7	19.6	0.534	0.43	0.89

R = remaining range of α-particle; the distance the α-particle would have traveled beyond the point of the fork if the collision were not incurred.
V = velocity of the α-particle in cm per second immediately before the collision.
Vp = velocity of the proton immediately after its escape from the heavy nucleus.
Rp = range of the proton, or H-particle.
Vn = velocity of the oxygen nucleus immediately after the escape of the proton.
Rn = range of the nucleus of the oxygen atom.
E_2/E_1= kinetic energy of the H-particle plus that of the oxygen atom, divided by the kinetic energy of the α-particle immediately before impact. 1-E_2/E_1 represents the loss of kinetic energy due to the nonelastic impact, and is presumably stored up in the oxygen nucleus of mass 17.
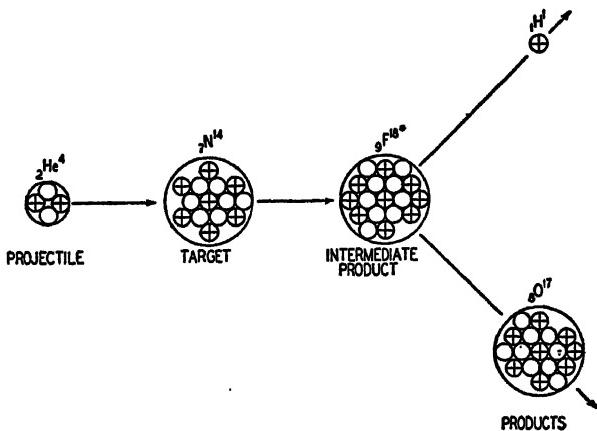
Figure 2. Formation of O^{17} and H^1 by the disintegration of F^{18}, the intermediate nucleus formed by the impact of He^4 with N^{14}. Open circles are neutrons; circles with plus signs, protons. At higher energies the excited fluorine nucleus disintegrates in two ways: (1) that given in the diagram, and (2) to form a neutron and positive electron in place of a proton (H^1). Inner circles represent only composition.
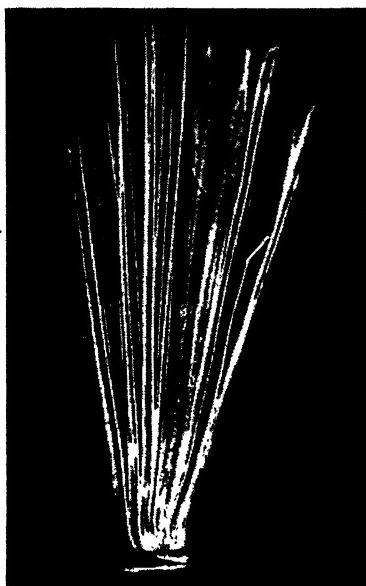
Figure 3. Same as Figure 1 except for energy relations.

in which a neutron and positive electron are emitted in place of the proton of equation (2).

It was pointed out by the writer that the relations of equations (2) and (4) give excellent evidence in favor of his theory of the intermediate nucleus. At low velocities of the helium nucleus (α-particle), the nitro-

gen nucleus is not penetrated, so that the first synthesis does not occur and a disintegration cannot result. As the velocity, and therefore the kinetic energy, of the α-particle increases, the synthesis begins to exhibit itself (Figure 9).

Although the existence of such an intermediate compound nucleus was, as stated previously, proposed by the writer in 1926, physicists did not accept the theory until ten years later. Their point of view was in general that expressed by Rutherford, Chadwick, and Ellis in 1930 as follows: [2] "It has been suggested by Chadwick and Gamow, partly on

Figure 4. Earliest photograph in which a thin sheet of solid was set across the chamber. *Below:* Tracks of α-particles; sheet used aluminum. *Above:* A proton gives a faint track.

$$_{13}Al^{27} + _{2}He^{4} \longrightarrow _{15}P^{31*} \longrightarrow _{14}Si^{30} + _{1}H^{1}$$
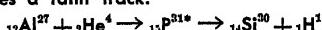
general grounds and partly on the basis of experiments by Chadwick, Constable, and Pollard, that the process of artificial disintegration of a nucleus by collision of an α-particle may take place in two ways: (1) by the capture of the α-particle by the atomic nucleus and the emission of a proton, (2) by the ejection of a proton without the capture of the α-particle. In the former case the α-particle must penetrate into the nuclear system, in the second case it seems likely that the disintegration would arise mainly from collisions in which the α-particle does not penetrate into the nucleus." This same point of view was also favored by Gamow in 1932.[3]

An apparent confirmation of the author's viewpoint came as the in-

direct result of a great discovery: that of the neutron by Chadwick [4] (described in a later section of this paper).

Why Noncapture Nuclear Disintegration Cannot Occur. It was obvious that if noncapture disintegrations occur, the theory of the inter-
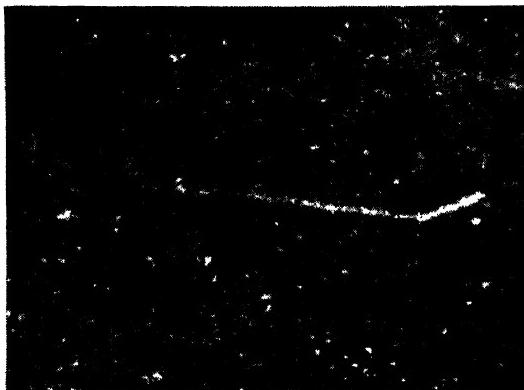
Figure 5. First neutron found in America. (Described in Figure 6.)

Figure 6. This figure gives the actual dimensions for the capture (Figure 5) of a neutron (dotted line) by a nitrogen nucleus of mass 14 (upper end of dotted line) to give a nitrogen atom of mass 15, which disintegrated into a B^{11} atom with a track 2.2 mm long and a He atom (track 12 mm long). The neutron had a velocity of 1.89×10^9 cm per second, and an energy of 3.00 million electron volts, and passed in a straight line through 9 mm of lead, 16 mm of glass and 64 mm of air. The velocity corresponds to 14,700 miles per second. This was the first neutron found in America. Source was Be + RaTh.
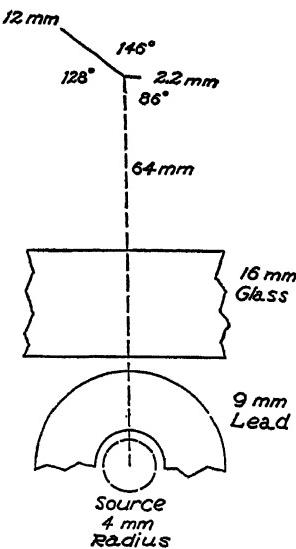
mediate nucleus must be invalid. Therefore it was essential to show that (1) the disintegration of nitrogen by noncapture of the neutron, as assumed by nuclear theorists, including those at Cambridge, does not occur, and (2) that the energy relations are unfavorable to disintegrations of this type. With reference to (1) it was shown that the disintegrations assumed to occur by noncapture actually occurred by capture. Feather,[5] who cooperated with Chadwick in obtaining Wilson cloud-

track photographs of collisions of neutrons with nitrogen, found two apparent types of disintegration: In less than one-half the cases the beryllium source of the neutrons was in the plane of the two tracks of the α-particle and the B^{11} nucleus which were produced. These he assumed to represent disintegrations by capture of the neutron. "The remaining cases" states Feather, "obviously represent disintegration without the capture of the neutron."
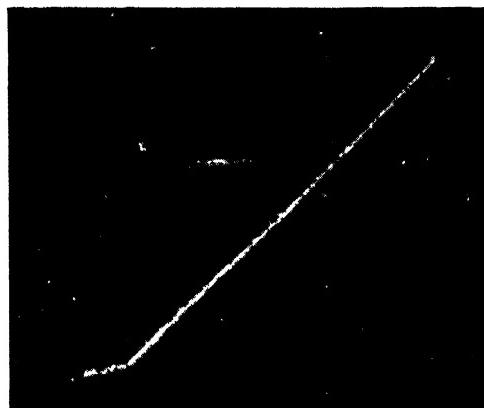
Figure 7. One of two almost identical photographs taken at right angles. This represents the disintegration of N^{14} by a neutron. Position of the source is such that if the neutron is not scattered, the disintegration must have been by noncapture, as assumed at Cambridge. However, actually it is due to the impact of a neutron which has been scattered in the apparatus and represents an event which occurred by capture.

Figures 5, 6, and 7 represent one of two views, at right angles, of tracks obtained with nitrogen. The long track is that of an α-particle and the short track that of a boron nucleus. By noncapture the reaction would be:

$$_7N^{14} + \text{k.e.} \longrightarrow {}_2He^4 + {}_5B^{10} + \text{k.e.} \qquad (5)$$

$$14.00751 \quad 0.003 \quad n \quad 4.00390 \quad 10.01618$$

$$\overline{\qquad\qquad 14.02008 \qquad}$$

$$\Delta m = 0.01257 + \Delta\text{k.e.}$$
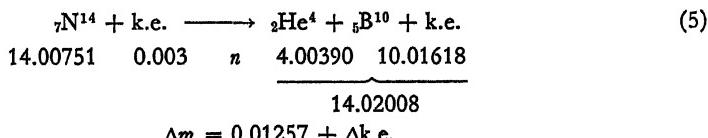
However, if the intermediate nucleus is involved:

$$_7N^{14} + {}_0n^1 + \text{k.e.} \longrightarrow {}_7N^{15*}$$

$$14.00751 \quad 1.00893 \quad 0.003 \qquad 15.01944$$

$$_8N^{15*} \longrightarrow {}_2He^4 + {}_5B^{11} + \text{k.e.}$$

$$15.01944 \qquad 4.00390 \quad 11.01284 \quad 0.00270$$

$$m = -0.00270$$
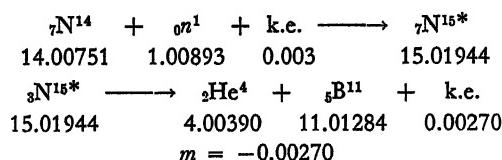
The numbers under the symbols represent the masses of the atoms. The kinetic energy of the neutron amounts to 0.003 unit of mass. Disintegration by *noncapture* [equation (5)] cannot occur since 0.01257 unit of mass (13.6 million electron volts of energy) would have to be *created*. If an intermediate nucleus is formed, the decrease in mass is 2.7 mg, which indicates that the reaction can proceed.

A further study of all of the reactions which were supposed at Cambridge to occur by noncapture disintegration, showed that they actually occur by capture. What caused the apparent discrepancy was that in
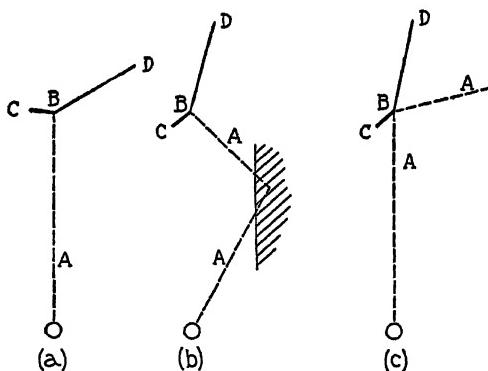
Figure 8. (a) Diagram of disintegration of nucleus *B* by capture of projectile *A* directly from source. (b) Diagram of the same disintegration by capture as for (a), except that projectile is first deflected by a nucleus in neighboring material. (c) Diagram of disintegration (which *never* occurs) of nucleus *B* by noncapture of projectile *A* directly from source.

every case of this type the neutron was *scattered* by a nucleus in the apparatus, before it united with a nitrogen nucleus to form the intermediate nucleus $_7N^{15}*$. Figure 8 compares diagrammatically the methods of disintegration.

A general formulation of the effect of a projectile *P* upon a target nucleus *T* to give an excited intermediate nucleus *I**, which later (say 10^{-17} second) disintegrates into *C* and *D*, is

$$T + P \longrightarrow I^* \longrightarrow C + D$$
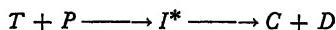
The following eleven paragraphs represent the state of the theory of the intermediate nucleus as expressed in 1934.

(1) The fundamental theory is that the intermediate nucleus has an actual life and a separate existence.

(2) The life period of any definite intermediate nucleus is independent of the components from which it has been formed, provided it is in a definite excited state, but the state of excitation depends upon both the kinetic and mass energies of these

components. For example C^{12*}, whether formed from $B^{10} + H^2$; $B^{11} + H^1$; or from $Li^6 + Li^6$, should have the same half-period, provided the sum of the mass energy and the kinetic energy is the same in all three cases.

(3) The life period depends upon the particular excited state of I^*, on the energy of this state, and upon the nature and the energy states of its possible products.

(4) When any definite intermediate nucleus is formed by the union of two definite nuclei the life period may in general be expected to decrease as the energy of the projectile increases. From the quantum viewpoint this is due to the general broadening of nuclear levels as the nuclear energy increases. On the classical basis it may be explained by the assumption that the greater the internal energy of the nucleus, the greater is the probability that a single particle (or group of particles such as $[(np)_2]$) should, in any stated interval of time, be given enough energy to allow it to escape from the nucleus. On this basis, the energy would be that of vibration, and the case becomes similar to a monomolecular reaction as considered in chemical kinetics.

(5) The intermediate nucleus I^* may disintegrate into the components T and P, as well as into other particles.

(6) On account of the low decimal mass of helium (0.0034 or 3.2 mev) the emission of helium is favored rather than of protons or neutrons, from the standpoint of the masses involved. Other relations and selection rules are considered later.

(7) According to the theory, the total number of disintegrations into nuclei must be equal to the number of syntheses of the excited intermediate nucleus minus the number of these nuclei which return to a stable state by γ-ray emission alone. Here disintegration into either heavy particles, or into negative and positive electrons is considered, but with light nuclei it is well known that heavy particles, such as neutrons, protons, α-particles, etc., are almost always emitted by the first or primary intermediate nucleus. An illustration of a case in which an electron is emitted, but not a heavy particle, is the following:

$$_6C^{12} + {}_1H^1 \longrightarrow {}_7N^{13*} \tag{6}$$

$$_7N^{13*} \longrightarrow {}_7N^{13*} + \gamma \tag{7}$$

$$_7N^{13*} \longrightarrow {}_6C^{13} + {}_1e^0 \tag{8}$$

Here, however, the primary excited intermediate product emits a part of its energy of excitation and changes into an unstable form. This intermediate, which has a half-life of 660 seconds, emits a positive electron.

(8) From (5), the total number of disintegrations is, in a secondary sense, dependent upon the nature of the constituents, T and P, including their energies, but is independent of the nature of the products, C and D, formed by the disintegration.

(9) The types of disintegration which occur, and their corresponding widths of level, depend upon the state of the intermediate nucleus, particularly upon its energy, and the angular momentum.

(10) The disintegration of the intermediate nucleus I^* may occur in steps, so that a second intermediate nucleus I'^* is formed.

(11) The lifetime of the intermediate nucleus is often 10^{-17} to 10^{-16} second.

More recently Wheeler [6] gives the life period for the intermediate nucleus Be^8 as 10^{-17} to 10^{-15} second for an energy of 125 kev and a width of from 1 to 100 ev. When γ-rays are emitted, the lifetime is 10^{-14} to

10^{-13} second. In general, the life period is of sufficient length to produce equipartition of energy through the nucleus. The long life period of the compound nucleus as compared with the characteristic time, i.e., the period if no compound nucleus were formed, which is about 10^{-22} second, is due to the fact that fluctuations, in which a large part of the energy becomes concentrated upon one nucleon or a small group of nucleons, are rare. In the case of fission, the two groups of nucleons are large instead of small. In considering the decay for a charged particle such as a neutron or proton, the potential barrier reduces the rate of decay, thus giving a longer life period.

Noncapture disintegrations do not occur because, unless the projectile penetrates the nucleus, not enough energy enters so that sufficient energy

Figure 9. The short track of this photograph represents the first radioactive nucleus ever produced artificially. It is a nitrogen nucleus of mass 16 which is definitely outside the range of nuclear stability and thus disintegrates, giving off a negative electron, into oxygen 16.

$$_9F^{19} + _0n^1 \rightarrow _9F^{20*} \rightarrow _2He^4 + _7N^{16} \qquad _7N^{16} \rightarrow _8O^{16} + \beta$$
$$\text{radioactive}$$

can concentrate itself on a single particle (or group) to cause its emission. From the standpoint of theory this may be true up to very high energies, possibly in some cases up to an energy as high as a billion electron volts (1 Bev) for the projectile.

Nitrogen 16—the First Artificially Produced Radioactive Nucleus. Immediately upon discovery of the neutron in 1932 by Chadwick it was assumed in this laboratory that if a fast neutron were to act upon a fluorine atom of mass 19, it would form a fluorine nucleus of mass 20 as the intermediate nucleus, which would then disintegrate into He^4 and N^{16}. The intermediate nucleus would have a ratio of neutrons to protons (N/P) of 1.286. Since this ratio is far above the range of stability in that region, further disintegration would occur; a negative electron would be given off to form a nucleus of oxygen of mass 16, in which the ratio is unity.

When this experiment was tried in a Wilson cloud chamber by filling the chamber with a gaseous fluorine compound, it was found that the above assumption was correct. In Figure 9 is shown a track of the first

radioactive nucleus ever produced artificially. This discovery was made in January, 1933, and was reported to a conference of the Century of Progress Exposition held in June, 1933.[7]

Evidence for the Existence of the Intermediate-Compound Nucleus. When, in 1934, the writer presented the above relations in the form of a paper to the *Physical Review* for publication, Bethe did not allow its publication on the ground that it is impossible for such an intermediate nucleus to have any existence whatsoever. There was much evidence that this statement is incorrect; there is no space here for all of the extensive evidence now available. It may be almost sufficient to say that at the present time no nuclear physicist, either theoretical or experimental, denies the existence of the intermediate nucleus.
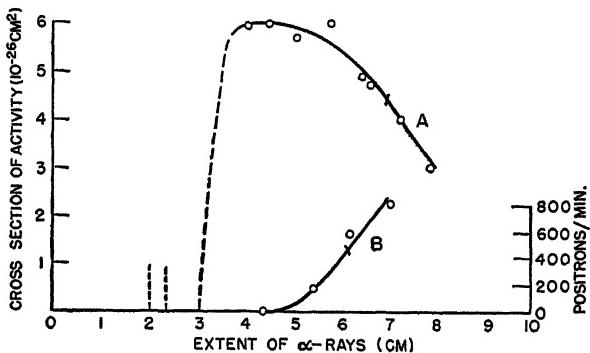
Figure 10. Excitation curve for the action of helium on nitrogen. Curve A gives proton emission; curve B, neutron and positive electron emission. (Work of Häxel.)

Possibly the first definite evidence of the existence of the intermediate nucleus was given in the early '30's by the work done at the University of Chicago laboratory on the disintegration of nitrogen by neutrons.[8] As cited earlier, it was proved that all of the disintegrations of nitrogen obtained by the action of fast neutrons involved the capture of the neutron and none of them were obtained by noncapture.

One of the best evidences for the intermediate nucleus is found in the impacts of high-speed α-particles on light nuclei, in which (above the energy at which disintegration does not occur) the total yield of disintegration processes is moderately independent of the velocity of the incident particles. This is especially indicated in cases where protons or neutrons are released as a result of collisions (Figure 10). In such cases it is found that the sum of the yields of protons and neutrons is quite constant over a large range of α-ray energy, even though the individual yields of protons and neutrons may vary greatly. Particularly in point is the effect of high-speed α-particles upon nitrogen nuclei,

as carried out by the use of the Wilson cloud chamber in the University of Chicago laboratory from 1923 to 1926. It was this investigation which gave rise to the idea that the α-particle is in every instance captured to give an intermediate fluorine nucleus of mass 18 and charge 9, i.e., the sum of the masses 14 plus 4 and the charges 7 plus 2. Later this investigation was extended to higher energies by Häxel.[9] The results are exhibited in Figure 10, where A represents the number of protons emitted, and B the number of neutrons emitted in a definite total number of events. From the theory of the formation of an intermediate compound nucleus, an increase in the number of neutron emissions (B) should decrease the number of proton emissions (A) by the same number. The figure shows that this prediction is valid.

When the action of fast protons on B^{11} was investigated at Cambridge in 1935,[10] the reaction was given as

$$_5B^{11} + {}_1H^1 \longrightarrow 3 \, {}_2He^4$$

The writer interpreted this as corresponding to the reaction series:

$$_5B^{11} + {}_1H^1 \longrightarrow {}_6C^{12}* \tag{9}$$

$$_6C^{12}* \longrightarrow {}_2He^4 + {}_4Be^8* \tag{10}$$

$$_4Be^8* \longrightarrow {}_2He^4 + {}_2He^4 \tag{11}$$

It was found later that the momenta of the two helium nuclei in reaction (11) were equal to each other when based on the position of the intermediate Be^8*, but that the momentum of the helium nucleus in equation (10) is not related definitely to the momenta of the other two. This shows that two independent reactions are involved and that the intermediate nucleus, $C^{12}*$, is actually formed. When a very high-energy proton is used, the lifetime of the Be^8* is exceedingly short.

The Intermediate-Compound Nucleus and the Fission of Uranium 236 or of Plutonium 240. In a nuclear transformation the first step is the formation of an excited nucleus, as follows:

$$P + T \longrightarrow C* \tag{12}$$

Then $C*$ must get rid of its excess energy by some type of change in which it loses energy, so that:

$$C* \longrightarrow D + E.$$
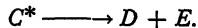
If E is excited, a second disintegration occurs (Figure 2). For the fission of uranium or of plutonium, the first step is the formation of the excited nucleus which is to split into parts. This is done by the addition of a neutron ($_0n^1$), as follows:

$$_{92}U^{235} + n^1 \longrightarrow {}_{92}U^{236}*$$

$$_{94}Pu^{239} + {}_0n^1 \longrightarrow {}_{94}Pu^{240}*$$
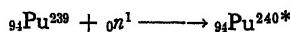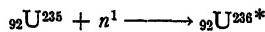
Since the disintegration by fission of these intermediate nuclei is of one type, only the fission of U^{236} will be considered here.

In its fission, the nucleus seems to split into two unequal parts of masses 127 to 154 for the heavier fragment and 115 to 83 for the lighter fragment. However, most of the masses lie between 134 to 144 for the heavier and between 90 and 100 for the lighter fragment. Inspection of Figure 12 shows that for U^{236} the ratio of neutrons to protons in the nucleus is 1.566, while in the range of mass from 134 to 144 even the highest ratios are in general 1.5 or less, and for mass numbers 90 to 100 the highest ratios of any of the isotopes are about 1.4. Thus, if the two atoms formed in the end as the result of any fission have the highest number of neutrons for stable isotopes, there are still neutrons to be accounted for.
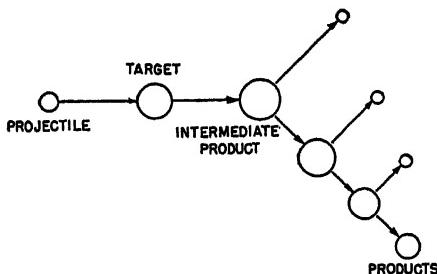
Figure 11. A nuclear transformation in which several intermediate nuclei are involved.

The process of fission may be considered analogous to what would occur if a charged water drop were to be set into strong vibration. This would result in the separation of the drop into two parts, but initially with a neck, or cylinder, of water between them. The neck then segregates into several minute water droplets. In the fission of U^{236}, it may be assumed by analogy, that the "droplets" formed are neutrons, 1 to 3 in number. The loss of this number is not sufficient to assure the stability of the two atoms formed as fragments. Therefore, these nuclei must expel neutrons, although the expulsion may be "delayed" for a period of moderate length. Another method of getting rid of extra neutrons is for the neutrons in the nuclei to lose negative electrons, thus converting neutrons in the nuclei into protons. It may be supposed that the "delayed" neutrons, observed to be given off after the fission occurs, are expelled after the emission of one or more negative electrons, since the life of atoms in electron emission is much greater in general than that for neutron emission. Although the water-drop model is used above to illustrate fission, nuclei in general seem to be more closely similar to minute crystals.

The history of the discovery of the fission of heavy atomic nuclei into two nearly equal parts is of extreme interest but cannot be related
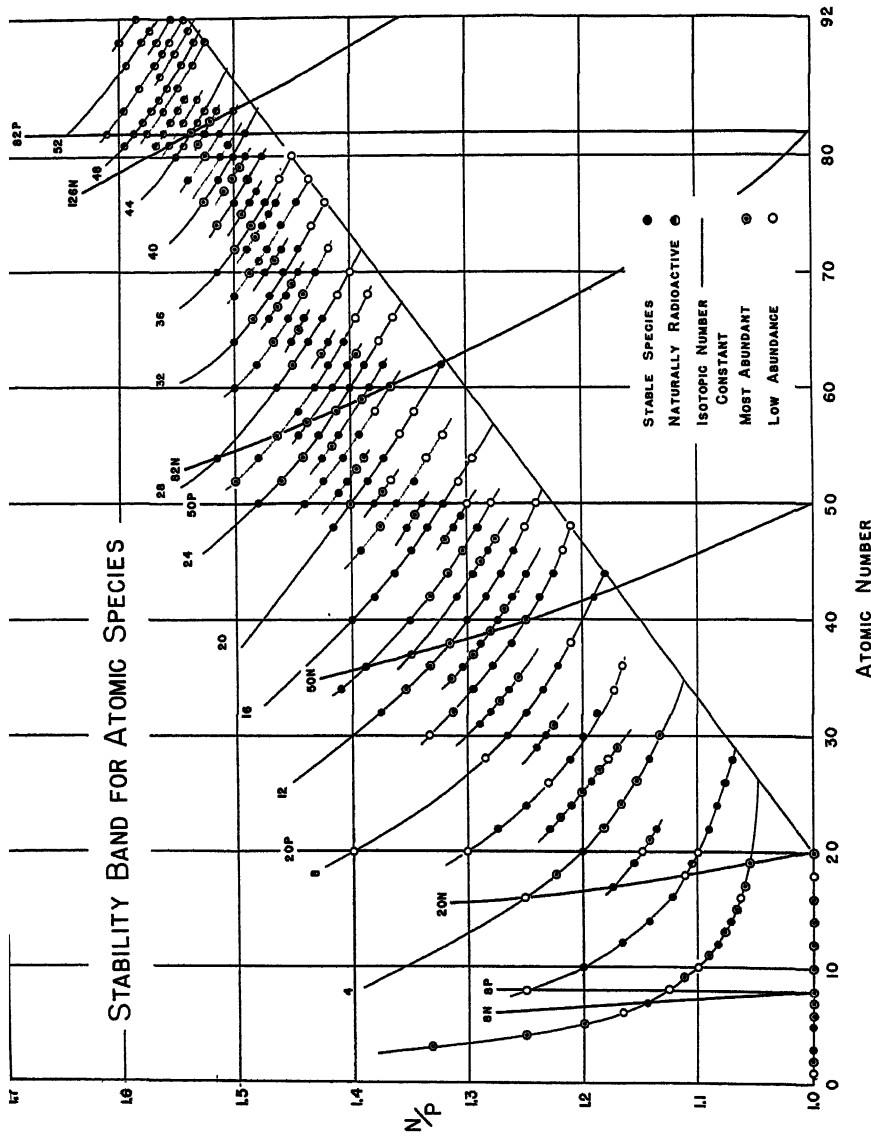
Figure 12. Ratio: (N/P) of neutrons to protons in atomic nuclei.

fully here. The lightest atoms such as the neutron, the hydrogen atom, or the atom of heavy hydrogen have masses higher than their respective whole numbers, in the case of the neutron by 0.9 per cent, and for hydrogen of unit mass by 0.8 per cent. The atomic weight of oxygen is a whole number by definition. As the atomic mass becomes higher, it falls below that of the corresponding whole number with extreme slowness; thus when barium of masses 130 to 138 is reached, the atomic mass has fallen to about 0.08 unit below the whole number. Above this mass range, the negative deviation from a whole number becomes smaller, so that when mercury of atomic weight 200 is reached, it is almost exactly at the whole number again (mass $200 = 200.016$). Above this, the positive departure from a whole number increases rapidly, and the mass of U^{236} may be given as 236.093.

If U^{236} were to split into two parts, it would lose the excess of mass over a whole number (0.093) and also 0.08, the defect in the mass, or 0.173, which is the sum of these. However, on the average, two neutrons are lost in the process, and these have an excess mass of 0.018, so that the loss of mass experienced by any 235 grams of uranium used in the bomb would be 0.155. This factor multiplied by c^2 gives 1.4×10^{20} ergs, or 3.4×10^{13} calories, which amounts to 2.7×10^{20} ergs, or 6.6×10^{13} calories per pound. The temperature thus given by the bomb is many million degrees and the pressure many millions of atmospheres. However, the efficiency is not perfect, since not all of the uranium or plutonium atoms in the bomb undergo fission, and the temperature is so high that a considerable fraction of the energy may escape as radiation, which, although it gives pressure, is relatively ineffective as compared with the motion of the molecules.

One of the most prominent properties of the neutron, due to its absence of charge, is that it can pass through an enormous number of other atoms without deflection, unless it strikes a nucleus. Since, on the average, about two neutrons are emitted by each fission and only one is needed to produce the fission of one nucleus, and since, in addition, the reaction is exceedingly fast, the number of neutrons and also the number of fissions increase with extreme rapidity, provided no neutrons escape. However, unless the body is moderately large (preferably spherical), so many neutrons escape that, on the average, less than one of the neutrons produced by a fission is effective in uniting with a nucleus to give another fission, and the reaction dies out. If the body is larger than this "critical volume" (highly dependent upon the shape), the reaction proceeds. The term, chain reaction, has been employed to describe the phenomenon, although this is somewhat different from the prior use of this expression in chemistry.

The Intermediate-Compound Nucleus and the Hydrogen Bomb. It has been decided to construct a hydrogen bomb, if this is possible. A group of scientists have stated that the particular isotope chosen is deuterium, which is hydrogen of mass 2. This decision is obviously based on two principal factors: (1) the isotope of hydrogen giving the greatest evolution of energy on reaction with itself (tritium) and (2) the most economically obtainable isotope, considering both cost and efficiency (deuterium). The choice of deuterium for the bomb is a compromise between these two factors.

In 1915, Harkins and Wilson made the first calculation of the energy developed by any nuclear reaction or group of reactions in order to determine whether the conversion of hydrogen into helium would give a sufficient amount of energy to heat the sun and the stars. They assumed that hydrogen, the most abundant element in the sun and stars, is converted into helium, the second most abundant, with a loss of mass of 0.03 gram per 4 grams of helium produced. By the Einstein relation the energy liberated would be $E = 0.03 \ c^2$ in which c, the velocity of light, is 3×10^{10}. This gave as the energy liberated 2.8×10^{10} ergs or 6.7×10^{11} calories for each 4 grams of helium produced. At that time astronomers did not favor this theory, since they considered that it did not give enough heat. If one pound of hydrogen reacts in the sun to form helium of mass 4 the amount of heat evolved is equal to that given by the combustion of 10,000 tons of coal. However, it was recognized that the hydrogen nucleus or proton is so exceedingly small (as previously stated, the radius of the proton is about 1.5×10^{-13} cm), that it would be impossible for more than two protons to react with each other in a sufficiently short time to give a nucleus heavier than a mass of 2 in a single reaction. Thus the only possible reaction would be that in which H^1 plus H^1 is converted into H^2 and a positive electron. This reaction does not give a sufficiently large evolution of energy.

Deuterium would be ideal for the bomb if it would react as follows:

$$H^2 + H^2 \longrightarrow He^4$$

to give a stable helium nucleus; but it does not do this. As mentioned earlier, in 1926 Harkins showed that all such nuclei are intermediate in the sense that they are excited and after an extremely short life disintegrate into two products. Actually, the reaction is:

$$H^2 + H^2 \longrightarrow He^{4*}$$

$$He^{4*} \longrightarrow He^3 + n + 4 \text{ mev}$$

or

$$\searrow H^3 + H + 4 \text{ mev}$$
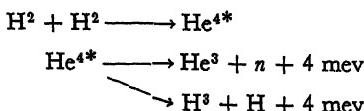
Thus only 4 mev is liberated from the intermediate He^{4*}. If tritium (H^3) were to be used, the reaction would be

$$H^3 + H^3 \longrightarrow He^6*$$
$$He^6* \longrightarrow He^5* + n$$
$$He^5* \longrightarrow He^4 + n$$
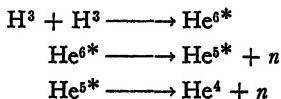
In this case two neutrons are liberated and the energy produced is 11.4 mev, so that tritium would give nearly three times as much energy as deuterium. If one tritium nucleus could be combined with one deuterium nucleus, one neutron would be given off and 17.6 mev would be liberated. The choice of deuterium is a compromise made on account of the fact that it is very much simpler to produce and very much less expensive than tritium.
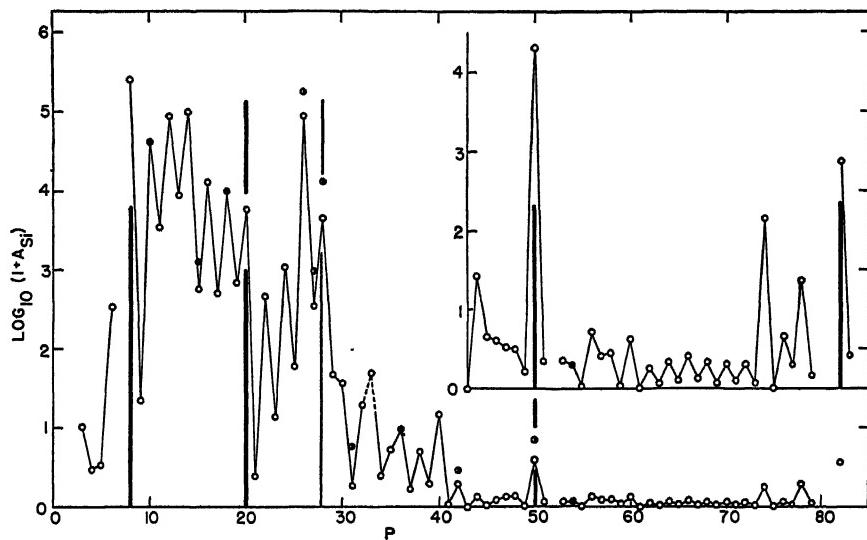
Figure 13. Abundance of the elements as a function of the number of protons (P), atomic number, in the nucleus. In almost every case, each even number of protons gives an extremely higher abundance than is exhibited for either of the adjacent odd numbers. Above P equals 43, two different sets of data were used, and one of these is represented on a magnified scale in the upper plot.

Nuclear Determination of Atomic Stability and Abundance

Stability of Atomic Nuclei and the Abundance of the Elements. In 1913, the writer began a statistical study of the atomic weights then known. The results indicated several relations of interest. The atomic weights (actually mean values for the element) were found to be so close to whole numbers for elements 2 to 28 inclusive that several conclusions seemed inevitable, as given in paragraphs (1) to (3) below. The general theory of nuclear structure advanced at that time by the writer was sufficiently fruitful to give rise to several other predictions, which were verified later [statements (4) to (6) below].

(1) The masses of the atoms are extremely close to whole numbers, with the exception of that of hydrogen (whole-number rule). This was confirmed by Aston in 1919 and later.

(2) Any considerable departure of the atomic weights (when correct) of the elements from a whole number was taken to indicate a mixture of isotopes. This conclusion was confirmed in the case of chlorine, atomic weight 35.46, by the separation of the element into isotopes.
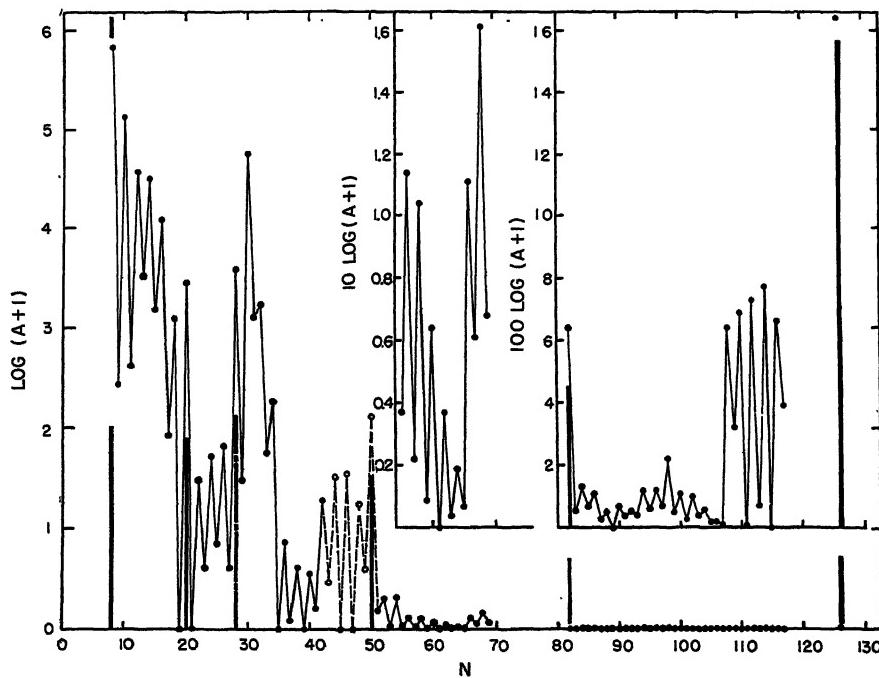
Figure 14. Abundance of the atomic species as a function of the number of neutrons in the nucleus. Almost universally each even number of neutrons gives a very much higher abundance than either of the adjacent odd numbers.

(3) It was found that for elements of even atomic number above 26 (even number of protons in the nucleus) the atomic weights are no closer to whole numbers than they should be by chance. Thus, for each of these elements there should be many isotopes, while the odd elements were indicated as having only one, or very few, isotopes. Later, this also was confirmed by Aston.

(4) The theory predicted that *elements* of even atomic number ($P =$ even) should be much more abundant than those of odd atomic number. When this was investigated, it was found that in the meteorites (Figure 13 gives similar relations for the universe) the elements of even

atomic number are seventy times more abundant than those of odd number, while in the surface of the earth and the sun this ratio is about ten to one. Also, the stars, in so far as they have been investigated, seem to exhibit a decided preference for elements of even number. Thus, atomic nuclei which contain an even number of protons are much more stable than those in which the number is odd.

(5) An equally prominent relation is that in almost all stable nuclei the number of neutrons is even (Figure 14). Thus, in so far as the abundance of atomic species has been determined, 99 per cent of all the atoms in the earth's crust have nuclei which contain an even number of neutrons. This relation became apparent to the writer in 1921, when he developed a formula for all nuclei as follows:

$$(pn)_z n_i$$

in which p is a proton; $n,$ a neutron; $z,$ the atomic number; and $i,$ the isotopic number. The isotopic number gives the excess in the number of neutrons above that for atoms whose nuclei contain equal numbers of protons and neutrons.

(6) Thus, there are almost no atoms with an odd number of both neutrons and protons whose nuclei are stable: only about one atom in ten thousand on earth. This is an extremely remarkable relationship (see Table 2).

TABLE 2. RELATIVE NUMBER OF NUCLEI PER ELEMENT

Elements	E–E	E–O	O–E	O–O
8 to 29	1,000,000	6,800	6,400	0
30 to 92	28	2.8	1.3	0

E = even nucleus   O = odd nucleus

The Equivalence of Protons and Neutrons. If the abundance of the atomic species in the universe is considered * with reference to the number of protons and of neutrons in the nucleus, a very remarkable relationship emerges. This relates the evenness or oddness of the number of protons or neutrons in nuclei (Table 2). In this figure the average number of nuclei per element from element 8 (oxygen) to 29 (cobalt) is arbitrarily placed at one million. These may be considered as even-even nuclei. If the number of protons is even and of neutrons odd, this value sinks to 6800, and the number of odd-even nuclei becomes 6400. These two numbers are equal within the limits of accuracy of our knowledge of the abundance of the atomic species. Thus it is found that it has been a matter of indifference in nature as to whether even-odd or odd-even nuclei are formed. Another extreme shrinkage in the

number occurs if we consider for these elements the number of odd-odd nuclei, since this value becomes zero; i.e., there are no atoms of elements 8 to 29 whose nuclei contain an odd number of protons with an odd number of neutrons.

Another remarkable relationship is exhibited when elements 30 to 92 are considered. Here the relative number of even-even nuclei is decreased from one million for the lighter elements to only 28. This indicates that in the building of atoms excessively more light than heavy atoms have been produced. The number of even-odd nuclei is now one-tenth of that of the even-even nuclei, since it is 2.8. The number for odd-even nuclei becomes 1.3. Whereas this is not exactly equal to the number of even-odd nuclei, nevertheless it is remarkably close to being equal if the inaccuracy in the data for these elements is taken into account. The number of odd-odd nuclei still remains zero.

If, with the elements 8 to 29, element 2 (helium) is included, then the number of even-even nuclei becomes very much larger and is in the neighborhood of 70 million instead of one million. No odd-odd nuclei occur except for elements 1, 3, 5, and 7, and of these the only abundant element is nitrogen, with 7 protons and 7 neutrons. No stable species of this type occurs with a higher number of protons and neutrons than 7. A few species which are extremely rare have been reported among the heavier nuclei, but in no case has it been shown that these species are stable.

The Existence of Stable Nuclei as Related to the Principle of Regularity and Continuity of Series and the Ends of Nuclear Shells. Recently it has been assumed by several theoretical nuclear physicists that groups of neutrons or of protons in the nucleus form what are considered as shells. A remarkable fact is that the shells end when a certain number of neutrons or protons are contained in the nucleus. The writer had already found that the numbers 2, 8 and 20 for either neutrons or protons give special stability to nuclei. The theoretical physicists involved in this important development found that certain nuclear properties revealed the values 2, 8, 20, 50 and 82 for both neutrons and protons and 126 for neutrons alone. This last value cannot apply to protons, since no nuclei contain as many as 126 protons. These particular numbers, which represent the ends of shells, have been designated by the physicists as *magic numbers*. In the Figures 15 to 20 these numbers for either protons or neutrons are represented by lines which are somewhat heavier than the others.

Although the writer assumed that in the building of the lightest nuclei protons were utilized, and for heavier elements neutrons, he found that stability relations of all atomic nuclei are essentially those exhibited by the naturally radioactive species which contain from 81 to 92 protons.
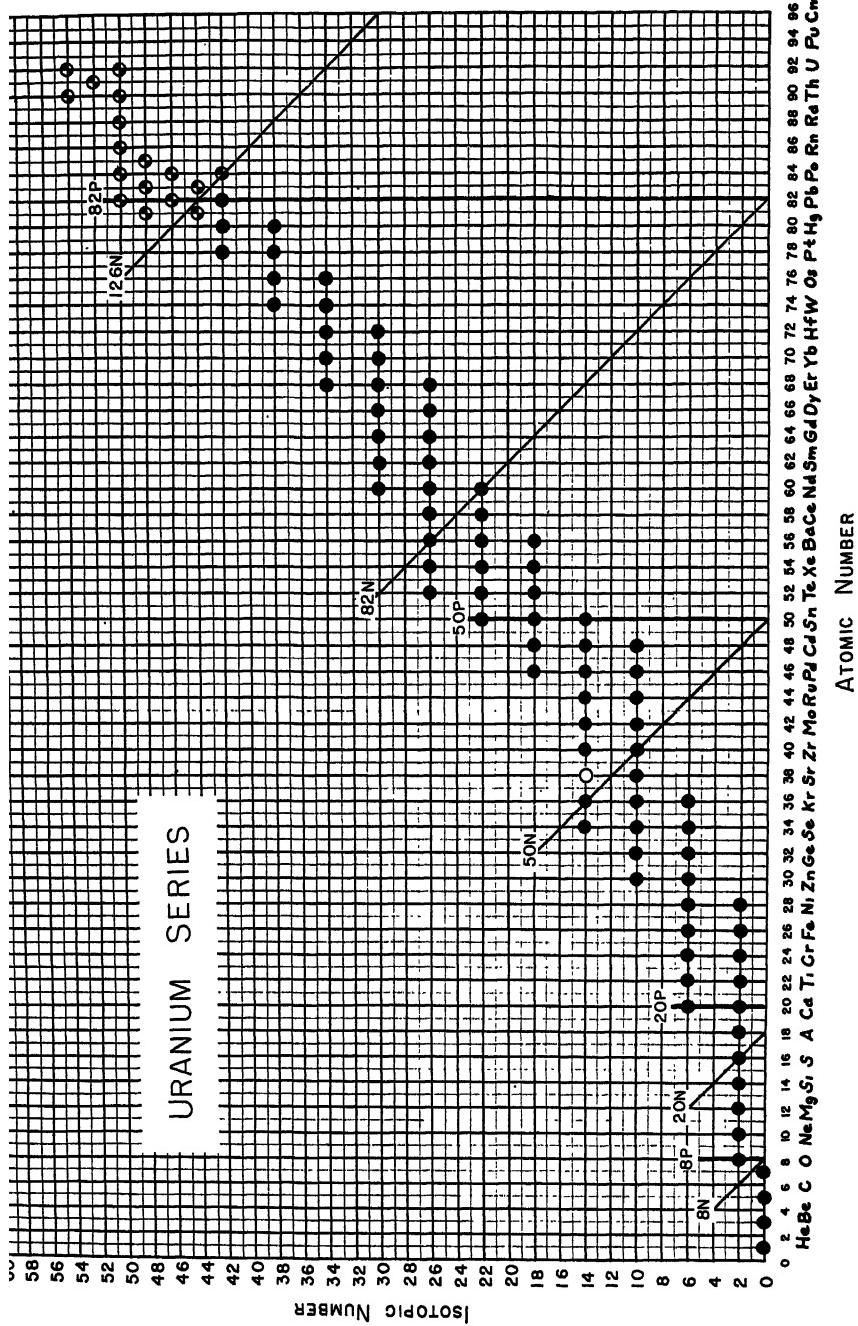
Figure 15. Uranium series. The ends of neutron and proton shells are indicated by heavier lines at values 8, 20, 50, 82, and 126. The one unstable species is strontium of isotopic number 14, designated by an open circle.

Figure 16. Helium-thorium series. Symbols are the same as in Figure 15.
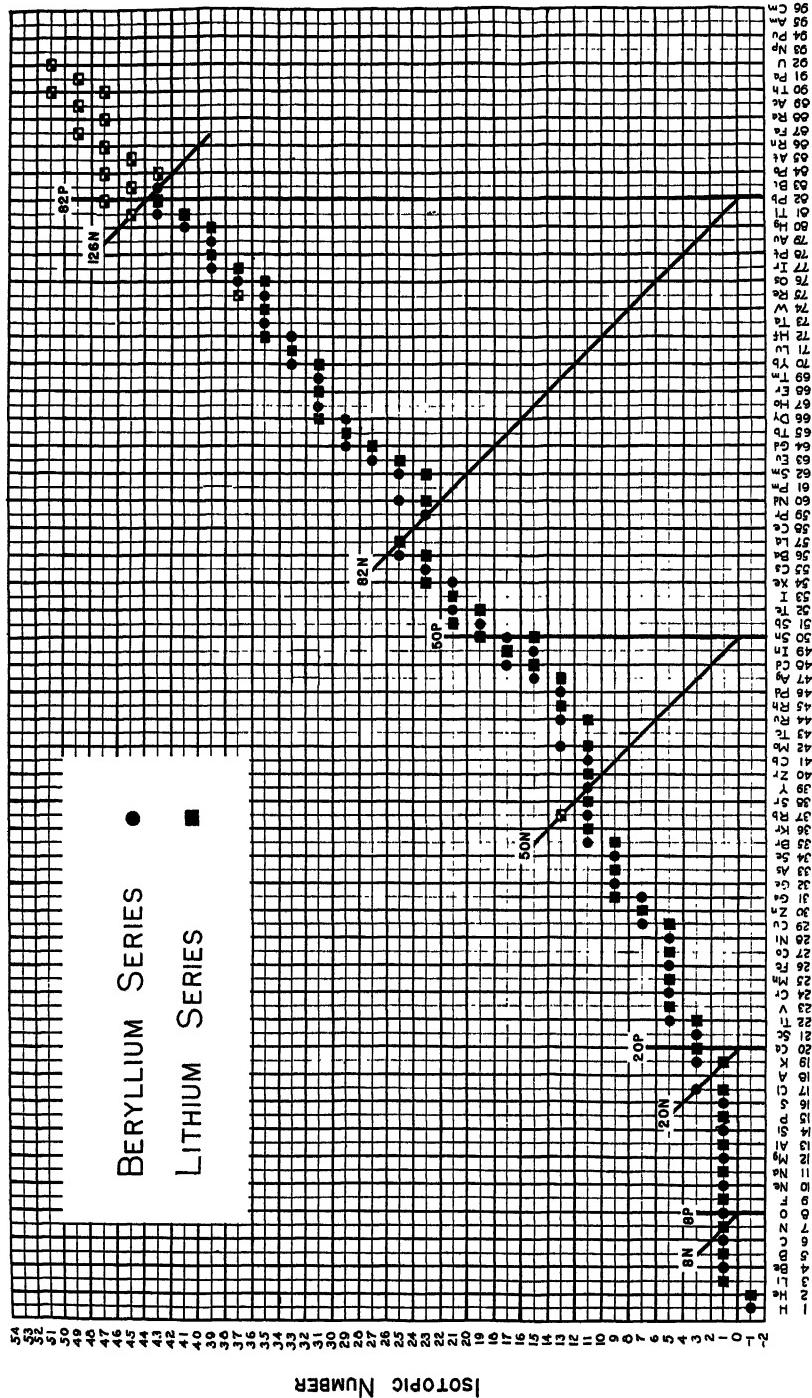
Figure 17. Stable species of the odd series. Heavy lines for 8, 20, 50, 82 neutrons or protons and for 126 neutrons represent ends of nuclear shells.
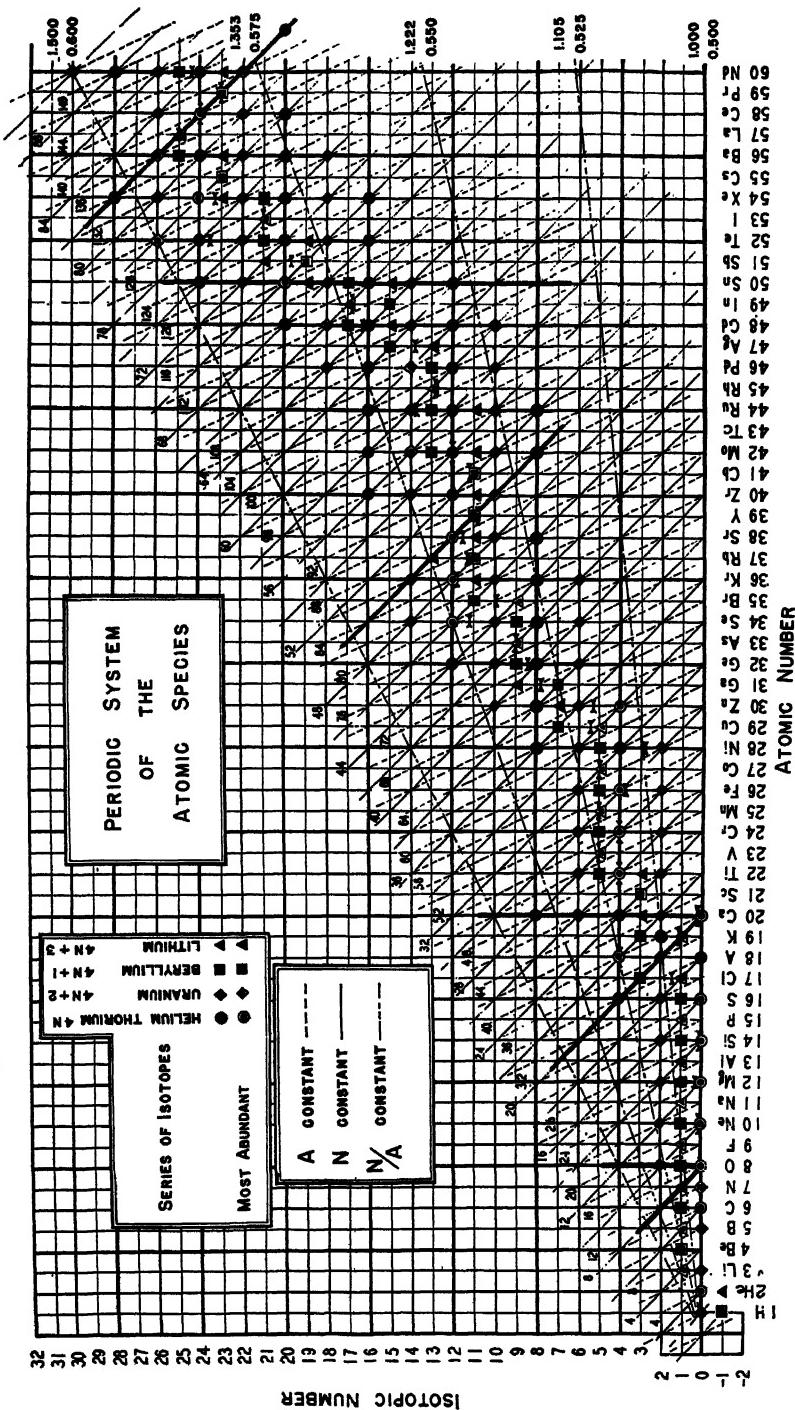
Figure 18. Stable atomic species to atomic number P equals 60. Heavy lines indicate ends of shells. N is neutron shell; P, proton shell.

It was found that the map which gives all of the stable species could be plotted in terms of three series: the deuterium-uranium, the thorium-helium (both even series), and the odd series, which is extremely different from the first two (Figures 15, 16, and 17). It is extremely remarkable that in the deuterium-uranium series (Figure 15) only one of all of the nuclei predicted by the principle of regularity and continuity of series is found to be unstable. All of the others are entirely stable. The relations exhibited in these three figures may be summarized as follows:

(1) No stable nuclear species is known *outside* the limits of these series in terms of proton (P) and neutron (N) content.

(2) No stable species is known *inside* the series, except those specified by the series.

(3) For even mass (M), both the number of protons and of neutrons is always even. No exception is known except for four species, in which $N = P$ with P very small (7 or less). For odd mass ($P + N$) either P or N may be odd, but not both. All other nuclei are excluded from the system of stable nuclei.

(4) Every species specified by a principle of continuity (considered later) exists and is stable, except in the relatively few cases in which some other principle intervenes and the nucleus is unstable [(5) and (6) below].

(5) For the even-even (N and P) series the stability and abundance are in general higher at the end of nuclear shells than for adjacent species not at the end of shells. The least stable and least abundant species are adjacent in proton and neutron content to the ends of shells. For example, in the uranium or $n + 2$ series, 83 positions are available, 82 are filled, and the one unfilled position is adjacent to the end of a shell.

(6) In the even-odd and odd-even series, unstable species are related to the non-occurrence of adjacent isobars. In these series 106 positions are filled, 10 are unfilled by the non-occurrence of adjacent isobars. Elements 42 and 61 are excluded, as are the odd isotopes of argon (see their peculiar pattern in Figure 17). Several of these are also close to the end of an 82-neutron shell.
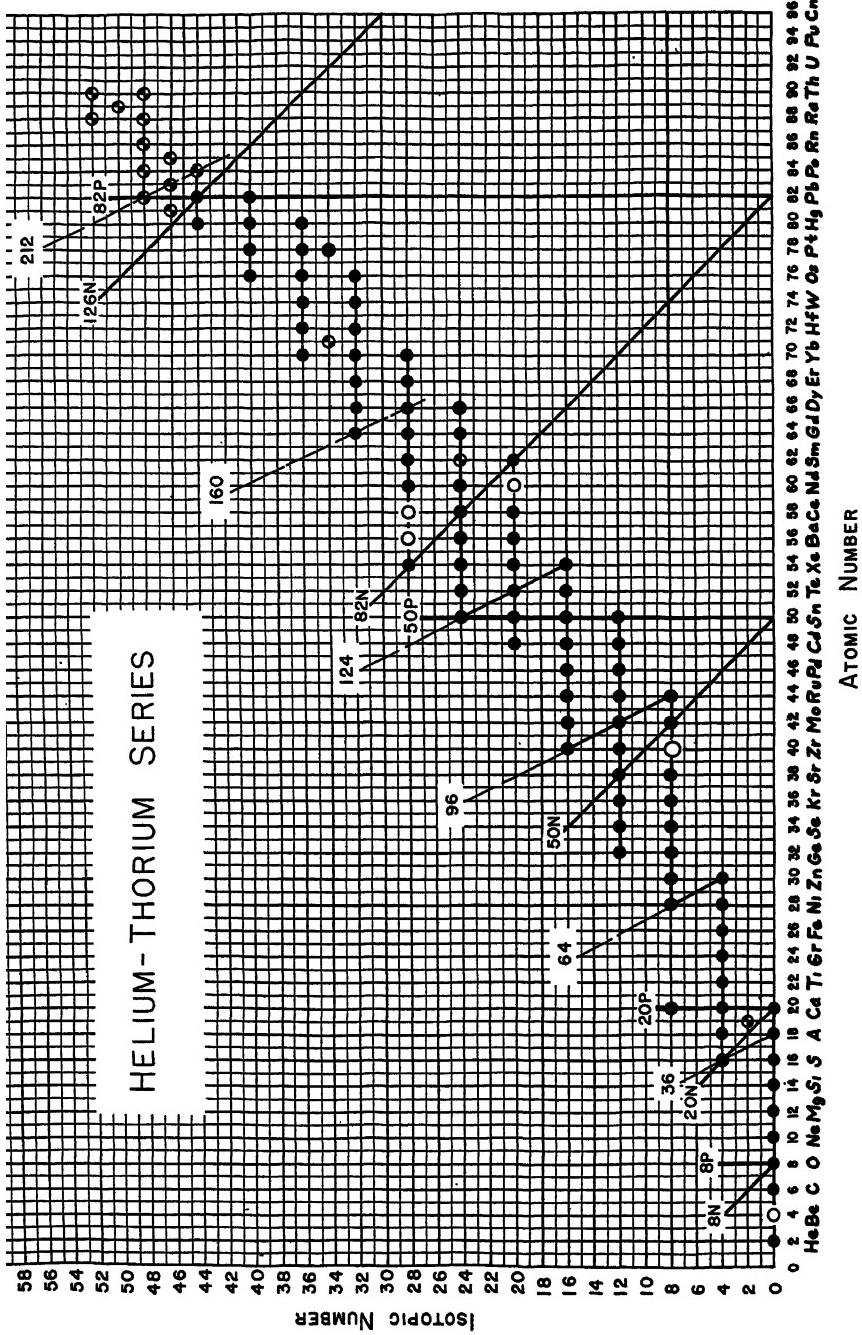
(7) In Figure 16 (helium-thorium series) it appears as though three atomic species predicted by the theory of continuity, of isotopic number 8, and with atomic numbers 22, 24, and 26, are missing as stable isotopes. However, these are not predicted by the theory, since the irregularity is due to the occurrence of the highest isotope of calcium of isotopic number 8, whose stability is related to the fact that it lies at the end of a 20-proton shell.

In Figures 18 and 19, all of the even and odd series are combined, and this combination reveals certain new relations as exhibited by the heavy lines. The figures give all of the stable species, also the naturally radioactive species of the U and Th series above $_{81}$Tl. The most abundant species are designated by a line around the symbol, and those below or slightly above 1 per cent of the element by an open circle. Above $P = 32$ these rare isotopes have the lowest neutron content, below this, often the highest. Very marked is the fact that for isotopes having a neutron MN of 50 (MN represents magic number) there are six species and three
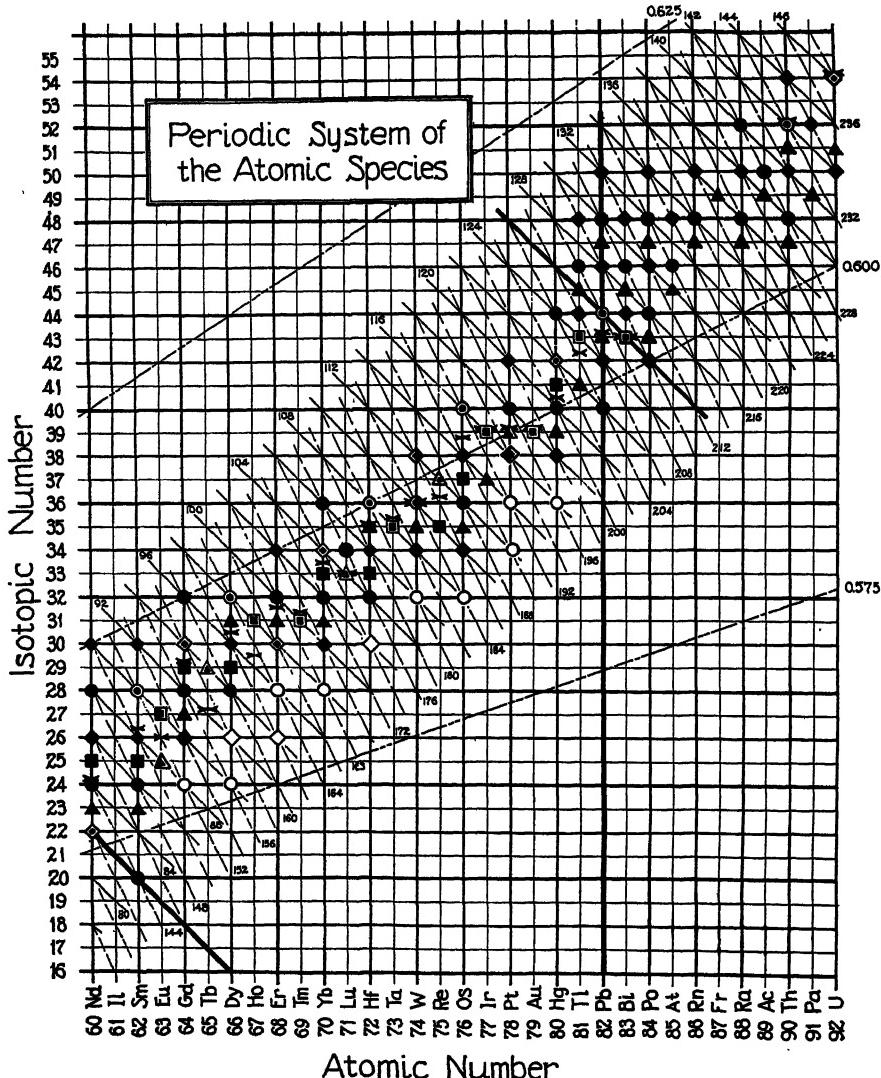
Figure 19. Stable atomic species. Atomic number P equals 60 to 81. Heavy lines indicate ends of shells. N is neutron shell; P, proton shell.

of these are the most abundant for the element. For neutron $MN = 82$, there are seven species, with five the most abundant for the element. The most marked effect of a proton MN is shown at calcium $= 20$, where

$$^{28}_{20}\text{Ca}^{48}_{8}$$ would presumably not exist if it were not at the end of a proton shell. This distorts the pattern. The region of few isotopes is just below this, but the neutron MN 20 gives four species, with two of them the

most abundant for the element. The 82-proton and 126-neutron magic numbers extend into the region of natural radioactivity, but even in this region the former is represented by three stable isotopes and the latter by two, and both of these two are the most abundant isotopes. .

A chart which emphasizes the effect of magic numbers on existence and isotopic number (extra neutrons) of the most abundant isotope is presented in Figure 20, where the values for the *odd* elements are connected by a line, and those for the even elements are not so connected. It is remarkable that for $P=9$ to 30, the number of extra neutrons (I) is represented by higher values for odd than for even elements, but for
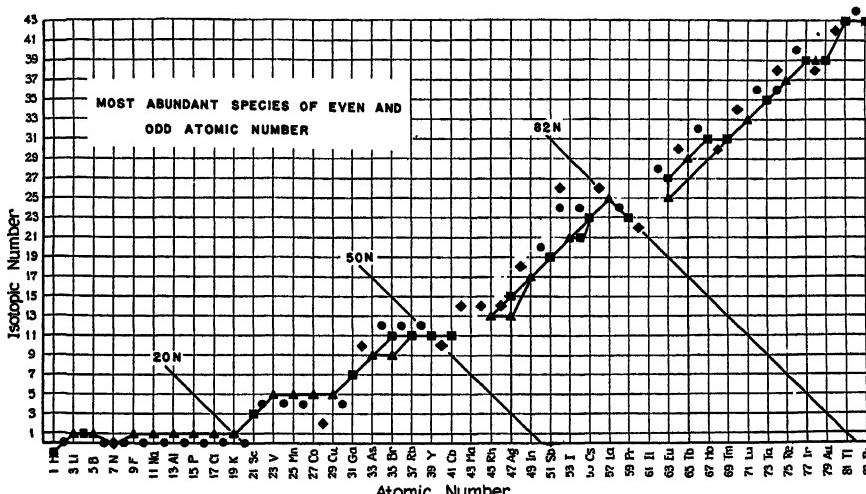
Figure 20. Most abundant isotopes of the elements of odd and of even atomic number. Heavy lines indicate ends of shells. N is neutron shell; P, proton shell.

32 to 77, the values of I for even elements lie above the line for odd elements. The shift occurs just where the region of abundant species changes to that of rare species. Just at Ni_{28}, where the shift occurs, the number of extra neutrons becomes specially low, i.e., two. The remarkable feature of the figure is the break which occurs just above the 50- and the 82-*neutron shells*. Here element 43, *which has no stable isotope,* lies just above the end of the 50-neutron shell, and 61, again with no stable isotope, just above the end of the 82-neutron shell. Neither the proton shells nor the 20- or 126-neutron shells have such a marked effect upon the relations discussed here.

In a paper submitted to the *Physical Review* in July, 1949, the abundance of nuclear species is considered in terms of even and odd P and even and odd N. This shows in a much more marked way the effects of the proton magic numbers. This method of treatment of even and odd P

was introduced in 1915 (1920 for even and odd N) by the writer, but prominent workers still classify $A = P + N$ according to even or odd number, which obscures certain of the relations, since odd + odd = even.

The effect of a proton magic number may be illustrated by tin at the ends of the 50-proton shell. Since the number of protons is greater than 30 it is a somewhat rare element. However, the effect of the end of the shell is shown by the fact that it is *much more abundant than any other element* from $P = 41$ to $P = 92$, and is, for example, twenty-four times more abundant than cadmium ($P = 48$), which has the next lower even atomic number.

During 1915 to 1917, the writer developed what has become known as Harkins' rule, that the elements of even number are very much more abundant that those of odd number. However, the remainder of the rule is usually forgotten; namely, that the atomic species which contain an even number of neutrons are very much more abundant than those in which the number is odd. Thus *2 is the most important of the magic numbers*. This is illustrated by helium p_2n_2 whose cosmic abundance, according to the latest estimate, is seventy times greater than that of *all* of the other elements, excluding hydrogen, which has a simple nucleus. This shell of 2 neutrons or of 2 protons has not received the attention which it merits.

For a discussion of closed shells with 20, 50, 82 neutrons or protons, or 126 neutrons, papers by Feenberg,[11] Maria Mayer[12] and Nordheim[13] may be consulted. Much earlier, J. H. Bartlett, Jr.[14] considered that closed shells exist in nuclei, and Gamow[15] in 1932 considered that Harkins had also introduced such shells. Actually, Harkins had pointed out the great importance of the special numbers 2, 8 and 20.

REFERENCES

1. Harkins, W. D., and Ryan, R. W., *J. Am. Chem. Soc.,* **45,** 2095 (1923).
2. Rutherford, Lord, Chadwick, J., and Ellis, "Radiations from Radioactive Substances," p. 572, New York, 1930.
3. Gamow, G., "Constitution of Atomic Nuclei and Radioactivity," Oxford, Clarendon Press, 1932.
4. Chadwick, J., *Proc. Roy. Soc., London,* **136A,** 692 (1932).
5. ———, *Proc. Roy. Soc., London,* **136A,** 709 (1932).
6. Wheeler, J. A., *Phys. Rev.,* **59,** 27 (1941).
7. Harkins, W. D., Gans, D. M., and Newson, H. W., *Phys. Rev.,* **44,** 945 (1933).
8. ———, ———, and ———, *Phys. Rev.,* **44,** 529 (1933).
9. Häxel, O., *Z. Physik.,* **93,** 400 (1935).
10. Oliphant, M. L. E., Kempton, A. E., and Rutherford, Lord, *Proc. Roy. Soc.,* A150, 241 (1935).
11. Feenberg, E., *Phys. Rev.,* **76,** 1275 (1949); and earlier papers.
12. Mayer, M. G., *Phys. Rev.,* **74,** 235 (1948).
13. Nordheim, L. W., *Phys. Rev.,* **75,** 1894 (1948).
14. Bartlett, J. H., Jr., *Phys. Rev.,* **41,** 370 (1932).
15. Gamow, G., "Constitution of Atomic Nuclei and Radioactivity," Oxford, Clarendon Press, 1932.

SCATTERING BY INHOMOGENEOUS MATERIALS *

P. Debye and A. M. Bueche

Dept. of Chemistry, Cornell University, Ithaca, N. Y.

Introduction

When a beam of light, or x-rays, traverses an inhomogeneous material its intensity is diminished. This diminution occurs partly because some of the radiant energy is transformed into heat and partly because another fraction of the incident light is scattered in directions other than that of the primary beam. Here we will only be interested in the scattered light. Most of it is of the same wavelength as the incident radiation. A small part, with which we will not be concerned, is of different wavelength and is known as Raman scattering.[1]

That scattering occurs at all is due to the fact that the materials illuminated are not perfectly homogeneous with respect to refractive index or, in the case of x-rays, electron density. In some pure substances these inhomogeneities are of atomic or molecular dimensions. In other materials, such as alloys and plastics, they may be much larger. Since the scattering depends on the characteristics of these inhomogeneities, its measurement provides a method for their study.

The interpretation of the scattering data which, in general, consist of the angular distribution of scattered intensity for different wavelengths and the total amount of radiation lost by the primary beam on traversing a known thickness of the sample, has been treated for a number of cases. It has been shown that the angular distribution and magnitude of scattering depend on the sizes and refractive indices of the inhomogeneities. Using the results of these treatments,[2-4] the method has been applied in investigations of many gases, liquids, crystals, and solutions. Usually the interpretations are made on the basis of some physical picture of the inhomogeneities, which may or

33

may not be a good approximation to the material under investigation, depending on its complexity.

There exist a number of materials, however, which are inhomogeneous and for which we have no clear-cut physical picture. The scattering is nevertheless still characteristic of the inhomogeneities. It is the purpose of this paper to elucidate a method[5] whereby data obtained from such a material may be interpreted so as to give information about the magnitude of the fluctuations in refractive index or electron density and the extensions in space over which these fluctuations occur. The use of the method will be illustrated by treating the data obtained when visible light is scattered by "Lucite." The authors feel the method will be useful in the study of a great variety of problems, such as the effect of inhomogeneities on the strength of materials, imperfections in crystals, and concentrated solution phenomena.

Characterization of the Inhomogeneities *

Consider an inhomogeneous material having an average dielectric constant ϵ, but with local variations η. These variations will be represented by a highly fluctuating function of the coordinates and will be both positive and negative so that the average value of η will be zero. In many cases it would be of value to know the magnitudes of these variations and the average distance over which they can be considered not to change appreciably. In what follows it will be shown how this information can be obtained in a large number of cases.

Two quantities are necessary to characterize these inhomogeneities. The first will be called the amplitude of η and is given by the average value of η^2 which will be indicated by $\bar{\eta}^2$. The second quantity, characterizing the average extension in space of the inhomogeneities must now be considered. Imagine two points, numbered 1 and 2, in the material, separated by a distance r. At point 1 there will be a fluctuation η_1 and at point 2 a fluctuation η_2. Form the product, $\eta_1\eta_2$, of these two fluctuations. Now move points 1 and 2 around in the material, keeping the distance r, between them fixed. In this way a large number of values of the product $\eta_1\eta_2$ can be found. The average value of the product will be indicated by $\overline{\eta_1\eta_2}$. This average will be a function of r. If $r = 0$, the value $\overline{\eta_1\eta_2} = \bar{\eta}^2$ will be found. For large values of r, the product will be zero because η_1 and η_2 will vary independently. In general, unless the material has a high degree of order so that the inhomogeneities are arranged in a regular fashion, $\overline{\eta_1\eta_2}$ will be a function of r, which goes from $\bar{\eta}^2$ for $r = 0$, to zero for large r in some, as yet, undetermined way.

To represent this variation with r, a new function, $\gamma(r)$, called the *correlation function*, will be defined by

$$\overline{\eta_1\eta_2} = \gamma(r)\,\bar{\eta}^2 \tag{1}$$

* A somewhat similar and independent treatment of the scattering problem herein discussed was attempted by C. L. Pekeris, *Phys. Rev.*, **71**, 268 (1947).

This function obviously must go from unity to zero for increasing values of r and characterizes the extension in space of the homogeneities.

Perhaps a very simple example of a correlation function would be illuminating. Consider a homogeneous liquid which contains a number of spherical particles of radius a, with a slightly different index of refraction than that of the liquid. In the limiting case of high dilution, in which interaction of the particles can be neglected, the correlation function is found to be

$$\gamma(r) = 1 - \frac{3}{4}\frac{r}{a} + \frac{1}{16}\frac{r^3}{a^3} \qquad\qquad 0 \leqq r \leqq 2a$$

$$\gamma(r) = 0 \qquad\qquad r \geqq 2a$$

This very special function starts with unity for $r = 0$ and becomes zero for $r = 2a$. At $r = a$ it has the value 0.312. In most cases the correlation function will not be nearly as simple as this. In each case, however, $\gamma(r)$ will be a measure of the extension in space of the inhomogeneities.

Scattering by the Inhomogeneities

Assuming η/ϵ to be small, it has been shown that the intensity scattered by a volume V will be proportional to the average square of the absolute value of the following integral, F, over the entire volume with element of volume $d\tau$.

$$F = \int \eta e^{ik\,(s,\,\vec{R})}\,d\tau$$

This integral is analogous to the amplitude of the form factor[4] of an atom for x-ray scattering. Here k stands for $2\pi/\lambda$ where λ is the wavelength measured in the medium, $\vec{R}$ is the vectorial distance from some fixed point whose position is immaterial, and $\vec{s}$ is a vector which is the difference between the unit vector, $\vec{S}$, in the direction of the scattered ray and that, $\vec{S_0}$, in the direction of the primary ray. That is,

$$\vec{s} = \vec{S} - \vec{S_0}$$

It follows that if θ is the angle between the primary and scattered rays, the magnitude of $\vec{s}$ is given by

$$s = 2 \sin \theta/2$$

The reader should be warned about applying the following method in cases where the preceding assumption of η/ϵ as small is not valid. It is probable that a few cases will arise, especially if visible light is used, where the quotient η/ϵ will not be sufficiently small. If the following method is used, the scattering data will lead to a correct picture of the material under investigation only if the primary beam on passing through the sample is not appreciably distorted by the presence of inhomogeneities. Although this problem [3, 6, 7] will not be

discussed in detail here, it may be well to remark that larger values of the quotient η/ϵ can be tolerated if the inhomogeneities are small than when they are large.

If the primary radiation is plane-polarized and the scattering is measured in a direction perpendicular to the direction of the electric vector, the scattered intensity will be given by

$$i \sim \iint \eta_1 \eta_2 \, e^{ik \, (\vec{s}, \, \vec{R_1} - \vec{R_2})} \, d\tau_1 \, d\tau_2 \tag{2}$$

In case the primary radiation is not polarized, a polarization factor $\dfrac{1 + \cos^2 \theta}{2}$

will have to be introduced. Remembering that $\vec{R_1} - \vec{R_2}$ is $\vec{r}$, connecting the points 1 and 2, and that $\overline{\eta_1 \eta_2}$ will be zero for distances r which are very small compared with the dimensions of the volume V, equation (2) reduces to

$$i \sim \bar{\eta}^2 V \int \gamma(r) \, e^{ik \, (\vec{s}, \, \vec{r})} \, d\tau$$

If, as in many cases, γ depends only on the distance r and not on its direction in space, the integral reduces to the final result

$$i \sim 4\pi \bar{\eta}^2 V \int_0^\infty r^2 \gamma(r) \, \frac{\sin ksr}{ksr} \, dr \tag{3}$$

However, if the medium is put under strain, or if the inhomogeneities are striations with a common axis, the scattered intensity may not have central symmetry around the direction of the primary ray. In such a case γ will depend on the direction of $\vec{r}$.

The next step is to investigate the properties of the final result in equation (3). Suppose the average extension of the inhomogeneities is so small that in the range in which $\gamma(r)$ is appreciably different from zero, $ksr = (4\pi r/\lambda) \sin \theta/2$ is small. Equation (3) then reduces to

$$i \sim 4\pi \bar{\eta}^2 V \int_0^\infty r^2 \gamma(r) \, dr$$

From this it can be seen that the scattered intensity is proportional to two quantities besides the illuminated volume V. They are $\bar{\eta}^2$ and a *correlation volume*

$$\omega_0 = \int_0^\infty \gamma(r) \, 4\pi r^2 \, dr$$

Thus only the product $\bar{\eta}^2 \omega_0$ is of importance. This leads to the very important result that, in this limiting case of inhomogeneities which are small compared to the wavelength, large fluctuations combined with small extensions are just as effective for scattering purposes as small fluctuations with large xtens icns.

In the more general case in which no assumption is made concerning the extension of the inhomogeneities as compared with the wavelength, the result can still be expressed in the same way. The scattered intensity is again

$$i \sim V\bar{\eta}^2\omega$$

that is, proportional to the product $\bar{\eta}^2\omega$, with the only difference that now the correlation volume ω depends on the direction of observation and the wavelength, since in general this volume has to be defined by the equation

$$\omega = 4\pi \int_0^\infty \frac{\sin ksr}{ksr} r^2 \gamma(r) \, dr \tag{4}$$

It is this dependence which enables us to draw conclusions about the extensions of the inhomogeneities independent of the value of $\bar{\eta}^2$.

So far, we have given only an expression, equation (3), which is proportional to the scattered intensity, without fixing the proportionality factor. Amplifying the statement in this respect, it can be shown that the scattered intensity i, coming from a volume V under an angle θ at a large enough distance R, is to the primary intensity I as

$$\frac{i}{I} = \frac{\bar{\eta}^2}{\epsilon^2} \frac{\pi^2 V}{\lambda^4 R^2} \frac{1 + \cos^2\theta}{2} \omega \tag{5}$$

in which ω is, as before, the correlation volume, and λ is the wavelength in the medium. In this formula the polarization effect is included by the introduction of the next to the last factor of equation (5).

By integrating over all directions in space, an expression for the turbidity τ can be derived from equation (5). In this way a formula is found, which shows a simple dependence on the wavelength only for the limiting case that the size of the correlation-volume is small compared with the wavelength. Under these circumstances, we have the familiar Rayleigh law that $\tau \sim 1/\lambda^4$ and we expect blue scattered from white incident light. From equation (5) it can be shown that for larger sizes of the correlation volume the scattered light will be more nearly white in color. In the limiting case of small size, the integration over all directions in space leads to the expression

$$\tau = \frac{8\pi^3}{3} \frac{\bar{\eta}^2\omega_0}{\epsilon^2\lambda^4} \tag{6}$$

for the turbidity τ in which the limiting correlation volume is defined by equation (4) and λ is the wavelength in the medium.

Determination of the Correlation Function $\gamma(r)$ from Angular Scattering Measurements

For the scattered intensity i, as compared with the primary intensity I, we have found

$$\frac{i}{I} = \frac{\bar{\eta}^2}{\epsilon^2} \frac{\pi^2 V}{\lambda^4 R^2} \frac{1 + \cos^2\theta}{2} \omega$$

in which the correlation volume ω is related to the correlation-function $\gamma(r)$ by equation (4). Instead of the angle θ itself, we can use the variable $s = 2 \sin \dfrac{\theta}{2}$ and plot the observed values of i as a function of this variable s. This plot being available, we can now plot the new function ϕ defined by

$$\phi = \left(\frac{s}{\lambda} \frac{i}{I}\right) \bigg/ \left(\frac{\pi}{4} \frac{V}{\lambda^4 R^2} \frac{1 + \cos^2 \theta}{2}\right) \tag{7}$$

which will have the dimensions of the square of a length, versus s/λ. In fact what has been done is to define this function ϕ by putting

$$\phi = \frac{2\bar{\eta}^2}{\epsilon^2} ks\omega = 4\pi \frac{\bar{\eta}^2}{\epsilon^2} \frac{s}{\lambda} \omega$$

This method of plotting is such, that in case experiments on the angular distribution have been made with different wavelengths, we will obtain just one curve for ϕ, provided it is plotted as a function of s/λ and not of s alone. Let us call the new variable σ, that is

$$\sigma = s/\lambda$$

ϕ, then, is a function of σ and as such can be found from the experimental data by the use of equation (7).

ϕ, however, is given by

$$\phi(\sigma) = 4\pi \frac{\bar{\eta}^2}{\epsilon^2} \sigma\omega = 8\pi \frac{\bar{\eta}^2}{\epsilon^2} \int_0^\infty r \gamma(r) \sin (2\pi\sigma r) dr$$

This integral equation can be solved at once by the well-known Fourier transformation and we obtain

$$2\pi \frac{\eta^2}{\epsilon^2} r \gamma(r) = \int_0^\infty \phi(\sigma) \sin (2\pi r\sigma) d\sigma \tag{8}$$

The only thing we have to do in order to determine the correlation function γ is to plot $\phi(\sigma)$, choose a value of r, plot the product $\phi(\sigma) \sin (2\pi r\sigma)$ versus σ, and determine the surface under this curve from $\sigma = 0$ to $\sigma = \infty$. This surface is now equal to the product $2\pi \dfrac{\bar{\eta}^2}{\epsilon^2} r \gamma(r)$. In this way we have found $\gamma(r)$ for the definite value of r which has been chosen. We can repeat the process for other values of r and so obtain the whole curve.

Application of the Method

An example of the application of this procedure to an actual case may help the reader. For this reason a description of the treatment of some data obtained by measuring the scattering of visible light by "Lucite" will be included.

The sample investigated was a piece of commercial "Lucite" rod. It was ground to the desired shape and polished. The scattered intensity was measured over a wide angular range, using both photographic and photo-

electric methods of detection. The data found using blue light, $\lambda = 4358\text{Å}$ of a mercury arc, are shown in Figure 1. The crosses refer to the data obtained photographically. On the ordinate are plotted the values of the scattered intensity, i, times $\sin \theta/2$, in arbitrary units after correction for the polarization by dividing the observed values by $\dfrac{1 + \cos^2 \theta}{2}$. On the abscissa are the values of $(\sin \theta/2)/\lambda$, which correspond to $\sigma/2$.

$(\text{SIN}\,\theta/2)/\lambda \cdot 10^{-4}$
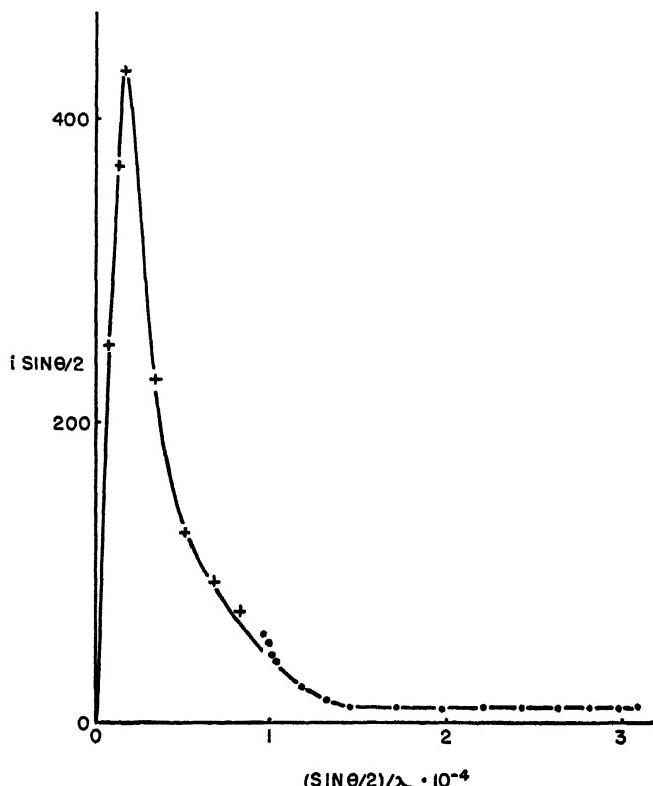
Figure 1. Scattering of blue light $(\lambda = 4358\text{Å})$ by "Lucite."

It will be noticed that the intensities did not fall off to zero for large values of $(\sin \theta/2)/\lambda$. This was taken to mean that, in addition to the large inhomogeneities responsible for the great angular dependence of scattering, there are a number of very small inhomogeneities which for this wavelength would give an almost constant background scattering. This background was subtracted from the data for the purpose of the following calculations. The results we obtain, therefore, will give us information about these inhomogeneities with the large extensions.
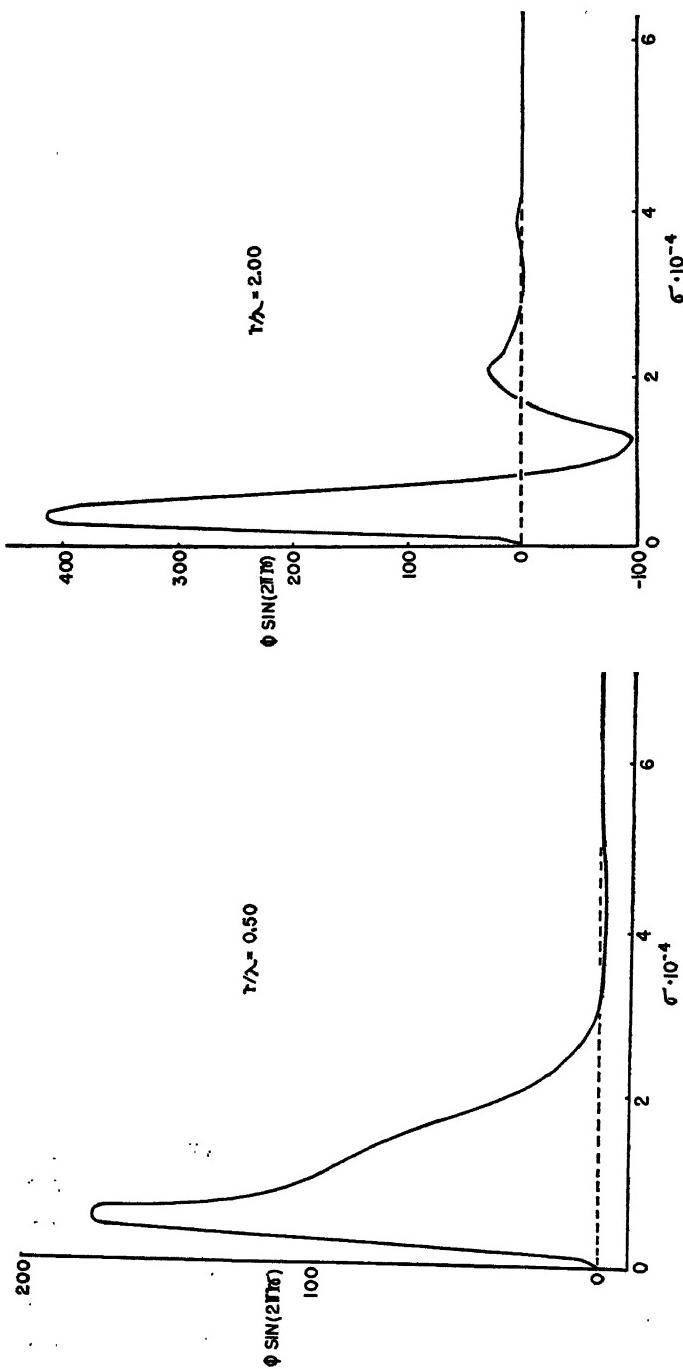
Figure 2. Examples of plots used to determine the correlation function.

From equation (7) it can be seen that for a single wavelength, and after correction for polarization, our $i \sin \theta/2$ is proportional to ϕ. It remains, then, to choose a value of r, multiply ϕ by $\sin (2\pi r\sigma)$, and determine the area under the curve of $\phi \sin (2\pi r\sigma)$ versus σ. From equation (8) it is obvious that this area is proportional to $r\gamma(r)$ for that particular value of r.

Figure 2 contains two curves of $\phi \sin (2\pi r\sigma)$ for two different r values in the case of "Lucite." $\phi(\sigma)$ is again in arbitrary units. In determining the total
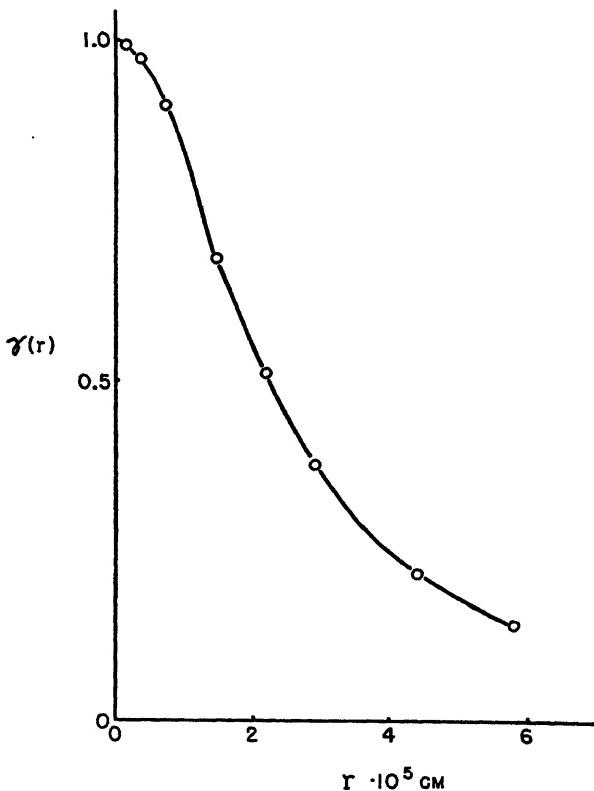
Figure 3. The correlation function, $\gamma(r)$.

area, that which lies below $\phi \sin (2\pi r\sigma) = 0$ must be subtracted from that which lies above.

By definition, $\gamma(r)$ is 1 for $r = 0$ and approaches 0 for large r. No matter what the units used, the points determining $\gamma(r)$ should be normalized so that this is true. Figure 3 is a plot of $\gamma(r)$ versus r for "Lucite." This function characterizes the extensions of the inhomogeneities.

For a comparison of different samples one could, for example, compare the r values of $\gamma(r) = 0.5$. In the case of "Lucite" this would correspond to

$r = 2250\text{Å}$. Another, perhaps more illuminating, way to treat $\gamma(r)$ is to try and fit a mathematical expression to the data. An example of such an expression is

$$\gamma(r) = e^{-r/a}$$

It is evident from Figure 4, where $1 + \log_{10} \gamma(r)$ versus r is plotted, that this is a fairly good approximation. In this case $a = 2800\text{Å}$. For studying different samples, values of a could be compared. The larger the a values, the larger are the extensions of the inhomogeneities.
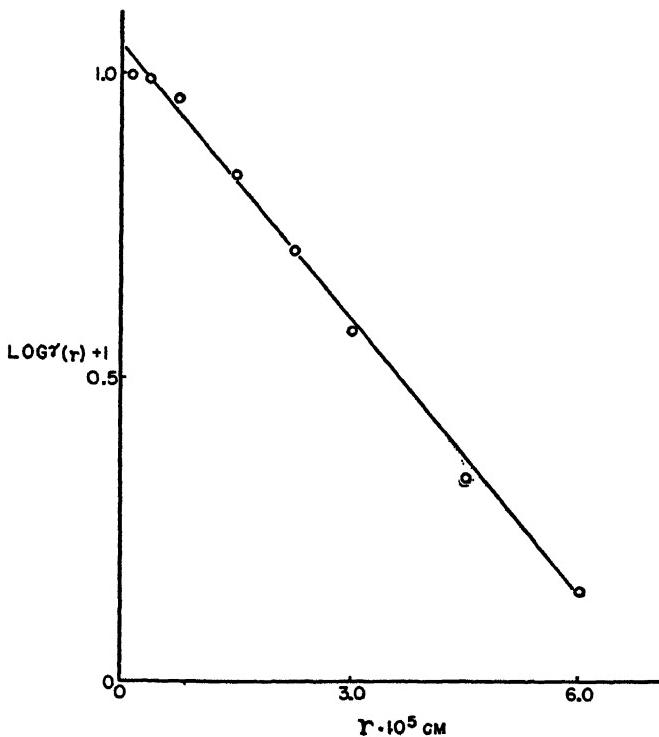
Figure 4. The correlation function approximated by $\gamma(r)$ equals $e^{-r/a}$.

Furthermore, if the exponential representation of $\gamma(r)$ is correct, the mathematics in the theory of the method become quite simple. It can be shown by substituting for $\gamma(r)$ in equation (3) and integrating that

$$i \sim V_{\bar{\eta}}^2 \frac{8\pi a^3}{(1 + k^2 s^2 a^2)^2}$$

Then, a plot of $\log\left(\dfrac{1}{i}\right)$ versus $\log\left(1 + k^2 s^2 a^2\right)$ should be a straight line.

Thus we have a convenient way of seeing how good the exponential approximation to $\gamma(r)$ is. In Figure 5 the logarithms of the experimental

intensities are plotted in the suggested way. The straight line is what would be expected if the exponential expression were a true representation of $\gamma(r)$. It seems to be a fairly good approximation.

The same type of plot can be used in another interesting and important way. If the angular dependence of the scattering from the solid is a result of interference of the scattered light, it should be possible to superimpose the data obtained using green light on that obtained using blue light by a way which considers the difference in wavelength alone. The data for green light, $\lambda = 5461\text{Å}$, have been plotted on the same scale as was used in Figure 5,
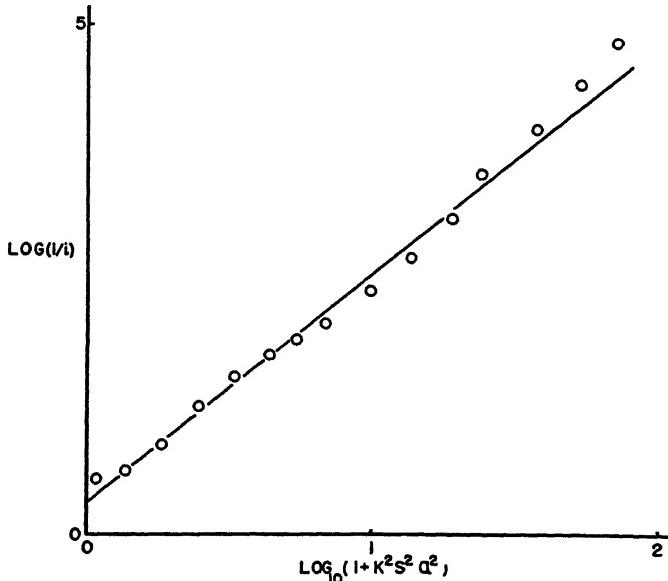
Figure 5. Comparison of the observed intensities (symbol o) with those calculated using $\gamma(r)$ equals $e^{-r/a}$ (symbol —).

the difference being only that the abscissae in this case have been calculated using a value of k corresponding to $2\pi/\lambda$ for green light instead of blue. This curve was then shifted, always keeping the axes of the two plots parallel, until a fit was obtained. The result is shown in Figure 6. The crosses refer to the points for green light, the circles are for the blue light. It can be seen that it was possible to superimpose the two curves. Therefore, we have evidence that we are really dealing with an interference phenomenon.

The formula for the turbidity in this case of the exponential correlation function can be shown to be

$$\tau = 32\pi^4 \frac{\bar{\eta}^2 a^3}{\epsilon^2 \lambda^4}\left[\frac{(b+2)^2}{b^2(b+1)} - \frac{2(b+2)}{b^3}\log{(b+1)}\right] \tag{9}$$

where

$$b = 4k^2a^2 = 16\pi^2a^2/\lambda^2$$

Knowing the extensions of the inhomogeneities, we would now like to know the average square, $\bar{\eta}^2$, of the fluctuations in dielectric constant which lead to this scattering. This can be done easily since the scattering at 90° of the "Lucite" has been compared with that of a standard solution whose turbidity, τ_0, is known. The intensity per unit solid angle, dI, of the light scattered by a solution of small molecules is given by

$$dI = I_0 \frac{3}{16\pi} \frac{dV}{R^2} \tau_0 (1 + \cos^2 \theta)$$
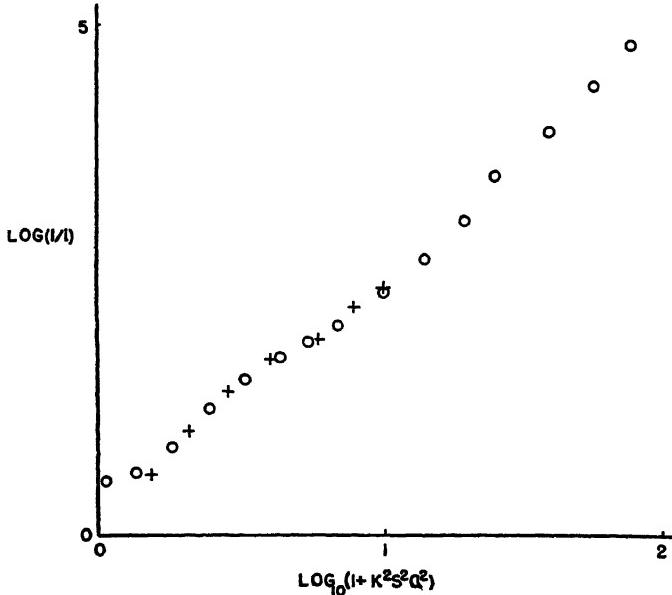
Figure 6. Superposition of the data for the Hg blue (symbol o) and green (symbol +) lights.

where dV is the scattering volume, R is the distance from the scattering volume to the point of observation and I_0 is the primary intensity. If dV/R^2 is the same for both solid and standard, the intensity scattered at 90° by the solid is

$$I = I_0 \frac{3dV}{16\pi R^2} \tau_0 \frac{p}{p_0}$$

where p and p_0 are the galvanometer deflections for the solid and standard, respectively. In the notation of the present work, $dV/R^2 = V/R^2$ and $dI = i$. The above expression can then be introduced into equation (7), and the absolute value of $\phi(\sigma)$ can be obtained.

Now, the surface of the $\phi(\sigma) \sin (2\pi r\sigma)$ versus σ curve for a chosen value of r is equal to $2\pi \frac{\bar{\eta}^2}{\epsilon^2} r\, \gamma(r)$, where ϵ is the average dielectric constant of the medium

and $\bar{\eta}^2$ is the average square of the dielectric constant fluctuations. Since we know $\phi(\sigma)$, and thus the surface, in terms of absolute values, we can pick a value of r, find the corresponding predetermined value of $\gamma(r)$, insert the proper value of ϵ^2, and find $\bar{\eta}^2$. In the case of "Lucite," using $\epsilon = 2.25$, it was found that $(\bar{\eta}^2)^{1/2} = 2.76 \times 10^{-4}$. This means that fluctuations in dielectric constant of the order of three parts in 10,000 cause the observed scattering of this sample. Since the dielectric constant is related to the refractive index, n, for long wavelengths by $\epsilon = n^2$ this corresponds to fluctuations in refractive index of about 1.7×10^{-4}.

Knowing $\bar{\eta}^2$ and the extensions of the inhomogeneities, we can find the turbidity of the "Lucite" sample. Inserting the proper values in equation (9) we find the turbidity to be 9.53×10^{-3} cm^{-1}. This is part of the total turbidity of the sample owing to the presence of large inhomogeneities. That caused by smaller fluctuations which we subtracted earlier is 9×10^{-4}, which is only one-tenth as large.

We may now compare these experimental values with the turbidity expected because of the changes in optical density owing to thermal fluctuations in a fluid. The turbidity, τ, due to these thermal fluctuations has been calculated by Einstein.[8] With the aid of the Lorentz-Lorenz equation it can be written as

$$\tau = \frac{8\pi^3}{27} \frac{kT}{\lambda^4} \beta [(n^2 - 1)(n^2 + 2)]^2$$

Here k is Boltzmann's constant, T is the absolute temperature, β is the compressibility of the fluid, λ is the wavelength of the light in vacuum, and n is the refractive index of the fluid. Using a compressibility $\beta = 3 \times 10^{-12}$ cm^2/dyne (an average value for glassy substances), and an index of refraction $n = 1.5$, we find the expected turbidity to be $\tau = 9 \times 10^{-6}$ cm^{-1}. The observed turbidity corresponding to the small background scattering was 100 times larger. This suggests that a study of this background scattering, using shorter wavelengths, may lead to information regarding molecular configurations in the solid.

Small Variation in Angular Scattering

It is clear that the procedure which has been discussed is only practical if and when the extension of the inhomogeneities is so large that a pronounced dissymmetry-effect exists. Otherwise one cannot cover with the given wavelength λ a large enough range of σ-values to define the function $\phi(\sigma)$ adequately.

We may expect that in our case of light-scattering the extension of the inhomogeneities may sometimes be too small in this sense. It is of importance to know what the conclusions are which can be drawn in such a case. In order to answer this question we can take our correlation volume as defined

by equation (4) and develop the function $(\sin ksr)/ksr$ under the integral in a power series. If this is done, we obtain

$$\omega = \omega_0 \left\{ 1 - \frac{k^2 s^2}{3!} \bar{r}^2 + \frac{k^4 s^4}{5!} \bar{r}^4 - + \cdots \right\}$$

in which we have used the following definitions

$$\omega_0 = \int_0^\infty \gamma(r) \, 4\pi r^2 \, dr$$

$$\bar{r}^2 = \frac{1}{\omega_0} \int_0^\infty r^2 \, \gamma(r) \, 4\pi r^2 \, dr$$

$$\bar{r}^4 = \frac{1}{\omega_0} \int_0^\infty r^4 \, \gamma(r) \, 4\pi r^2 \, dr$$

The scattered intensity i divided by the primary intensity finally becomes

$$\frac{i}{I} = \frac{\bar{\eta}^2}{\epsilon^2} \frac{\pi V \omega_0}{\lambda^4 R^2} \frac{1 + \cos^2 \theta}{2} \left\{ 1 - \frac{k^2 \bar{r}^2 s^2}{3!} + \frac{k^4 \bar{r}^4 s^4}{5!} - + \cdots \right\} \qquad (10)$$

We can always represent our experimental results by a series of the form

$$\frac{i}{I} = A \frac{1 + \cos^2 \theta}{2} \left\{ 1 - \alpha s^2 + \beta s^4 - + \cdots \right\} \qquad (11)$$

Comparison of equation (10) and equation (11) shows that the absolute intensity of scattering, measured by A, determines the product $\bar{\eta}^2 \omega_0$. As soon as an angular dissymmetry can be observed, the coefficients α, β, . . . can be evaluated. Perhaps only the first of them, α, if the dissymmetry is slight; more of these coefficients if it is more pronounced. If α alone is available, we only learn the value of $\bar{r}^2$ as a measure of the extension of the inhomogeneities. However, the more such coefficients are available, the more we learn about details of the correlation function $\gamma(r)$. The problem, how to make approximations for $\gamma(r)$ if $\bar{r}^2$, $\bar{r}^4$, etc. are known, can be handled mathematically. At the present stage of the whole problem we do not think it necessary to discuss this aspect. It may be sufficient to say that in the case of the distribution curve for the molecular weights of a polymer just the same problem occurs.

REFERENCES

1. Bhagavantam, S., "Scattering of Light and the Raman Effect," Brooklyn, Chemical Publishing Co., 1942.
2. Debye, P., *J. Phys. Coll. Chem.*, **51**, 18 (1947).
3. Oster, G., *Chem. Revs.*, **43**, 319 (1948).
4. Zachariasen, W. H., "Theory of X-ray Diffraction in Crystals," New York, John Wiley & Sons, Inc., 1945.
5. Debye, P., and Bueche, A. M., *J. Applied Phys.*, **20**, 518 (1949).
6. Mie, G., *Ann. Physik*, **25**, 377 (1908).
7. Stratton, J. A., "Electromagnetic Theory," New York, McGraw-Hill Book Co., Inc., 1941.
8. Einstein, A., *Ann. Physik*, **33**, 1275 (1910); trans. in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. I, New York, Reinhold Publishing Corp., 1926.

THE ROLE OF THE ELECTRIC DOUBLE LAYER
IN THE BEHAVIOR OF LYOPHOBIC COLLOIDS

E. J. W. Verwey

Philips Research Laboratories, Eindhoven, Netherlands

LYOPHOBIC COLLOIDS ARE characterized by the fact that their stability relations (peptization and coagulation), their rheological properties, their light-scattering and other physical properties are mainly determined by the electric charge of the particles, or better, the electric double layer surrounding the particles. According to this definition the concept of lyophobic colloids includes classical colloids, such as the gold sol, the arsenic-sulfide sol, the iron-oxide sol, the silver-iodide sol, etc., as well as the technically important groups, e.g., suspensions, clay systems, various types of emulsions, and certain protein solutions.

To understand the fundamental properties of such systems, it is best to consider two types of problems separately: (1) Problems concerned with the electrical conditions at the interface, such as the origin of the double layer, the extension of the double layer, electric charge, and potential distribution in the neighborhood of the interface, etc. (2) Problems concerned with the interaction of two or more particles. It is now generally accepted that this interaction is determined by at least two types of forces: van der Waals-London attraction forces, and forces due to the interaction of the double layers. In this paper a brief review of a few aspects of our present knowledge as to these two points is given.

(1) The Electric Double Layer of a Single Particle

The question of the origin of the double layer is still unsettled in a number of cases, but in most sols we have a pretty clear picture as to the nature of this double layer. We may distinguish between two types of double layers, which will be treated separately under (A) and (B) below.

Type (A). In a number of sols the double layer is predominantly due

to adsorbed capillary-active ions. A typical example is an oil-in-water emulsion emulsified by the addition of a soap.

At the interface of the oil and the water, in the absence of the emulsifier, there is already an electric double layer, caused by the fact that small amounts of electrolytes, which are always present in the system, generally tend to distribute their positive and negative ions unequally between both phases. Such a double layer however, is not sufficient to stabilize the emulsion.[1] It can easily be shown that the electric-potential difference corresponding to this double layer is dependent only on the nature of the electrolyte and is independent of the concentration of the latter. Accordingly, this double layer cannot be strengthened by a change of the electrolyte concentration. In addition, however, there is the important point that this double layer is of the double diffuse type. This means that both charge layers extend over a certain distance into the corresponding phase. Especially in the nonaqueous phase (or more generally in the liquid with the lowest ability to dissolve electrolytes and with the lowest electrolyte-dissociating power), the charge is spread far into the solution and the charge density in this part of the double layer accordingly is very low. The double-layer capacity of such a double layer at the interface of two immiscible liquids is extremely low, and we shall see later that this means that one of the conditions favoring stability of the sol is *not* fulfilled.

Relationships are altered, however, if a small amount of an electrolyte of such a nature that either the cation or the anion is highly capillary-active is added to the system. In this case a monomolecular layer of this ionic species is accumulated at the interface and a double layer of a different type is formed. Thus, one charge layer is built up by a thin, adsorbed ionic layer which, as a first approximation, can be considered a simple surface charge. This surface charge is generally very much larger per unit area of the interface than the charge of the original double layer. Accordingly this surface charge is balanced at both sides of the interface by diffuse charge layers, each of the same sign, but opposite to the sign of the surface charge. It can easily be shown [1a] that by far the largest fraction of this counter charge will be present in the diffuse layer in the aqueous phase, or more generally in the phase with the largest electrolyte-dissociating power. Hence as a consequence of the addition of the capillary-active electrolyte, the charges are redistributed in such a way that the original double layer is replaced by a triple layer. For all practical purposes, however, this triple layer can be considered a simple double layer consisting of a surface charge balanced by a diffuse charge in the aqueous phase with a much larger double-layer capacitance than the original double layer.

The simplest theory for the distribution of the charge and for the

electric-potential function in a double layer has been given by Gouy,[2] and later independently by Chapman.[2] In this theory the liquid charge, carried by ions moving freely in the solution, is treated as a continuous charge. The distribution of these ions is a compromise between the electrical attraction force due to the oppositely charged surface charge and the thermal motion which tends to move the ions farther into the solution. The theory shows that the charge density in the solution decreases rapidly as one goes farther from the surface and also the electric potential declines more or less exponentially as a function of the distance from the surface. Most of the charge falls within a plane at a distance $1/\kappa$ from the surface, in which κ is determined by

$$\kappa^2 = \frac{8\pi n\, v^2 e^2}{\epsilon k T}$$

In this expression, v is the valency of the ions, with a sign opposite to that of the surface charge, n is the number of these ions per cm³ in the solution, far from the double layer, e is the elementary charge, and ϵ is the dielectric constant of the solvent. Accordingly, the "thickness" of the double layer is proportional to $n^{-\frac{1}{2}}$. In the foregoing, we have already mentioned that in a medium with a low ionic concentration the liquid charge extends farther into the solution than in the case of a medium with higher ionic concentration. The expression also shows that the thickness of the double layer is inversely proportional to the valency of the counter ions. Hence the electrolyte added to the system reduces the double-layer thickness, and this effect is more pronounced the higher the valency of the counter ion.

A decrease in the double-layer thickness is equivalent to an increase in the capacitance of the double layer. As the charge of the double layer is determined by the amount of capillary-active ions in the interface and will be, roughly, a constant independent of the electrolyte concentration in the aqueous phase, we may infer that an increase in the electrolyte concentration must necessarily change the electric-potential function in the solution in such a way that the total potential drop in the aqueous charge layer (or the double-layer potential) is lowered.

Type (B). It is well known that in many lyophobic colloid systems the electric double layer is certainly not due to the presence of an adequate quantity of capillary-active ions. In the case of these systems, actually the more usual ones, the origin of the double layer has long been a more or less open question, although all sorts of qualitative considerations have been given, but for the most part relating to a special colloid. A more general solution of this problem has been attained as a result of the work of Lange [3] and his collaborators. In a very careful quantitative study of the electric double layer of the silver halides in

contact with water, Lange introduced the concept of potential-deter-
mining ions, hereby placing the question of the build-up of the electric
double layer on the sound basis of very simple thermodynamic
principles. Lange's argument is in principle involved in earlier work on
the double layer (e.g., Frumkin's investigations on the double-layer
metal/electrolyte solution). However, Lange extended the basic ideas
of Nernst and van Laar, so that the importance of these considerations
to problems in lyophobic colloids became much more obvious.

In later work by the author and other investigators from Kruyt's
school, Lange's views proved to be very useful in the study of lyophobic
sols of the silver-halide type.[4] Furthermore, his concepts could easily be
made directly applicable to all sorts of lyophobic systems (for instance
the oxide sols).

If a solid comes into contact with a polar liquid and ionic equilibrium
has been established, it is generally possible to indicate a pair of ions
(+ and −) in the solution, which are responsible for the double layer
formed at the interface and determine the electric-potential drop be-
tween the two phases. They are the ions which are contained by both
phases in common (if we may consider the solid as a substance built up
by ions), or which are involved in the electrolytic reaction at the inter-
face establishing the ionic equilibrium. One example is AgI in contact
with a diluted-aqueous solution of electrolytes saturated with AgI. Here
clearly Ag^+ and I^- are the potential-determining ions. Also Br^- or Cl^-
may, if present in the system, act to some extent as potential-determining
ions, because these ions can be built into the AgI lattice and may accord-
ingly influence the double-layer potential at the surface of the particles.
Another example is Al_2O_3 in contact with water or an aqueous elec-
trolyte solution. Here the O^{2-} ions in the surface are involved in the
reaction

$$O^{--}_{surface} + H_2O \rightleftharpoons 2OH^-_{solution}$$
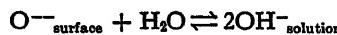
and the surface charge will therefore be determined by the OH^- (and
accordingly also the H^+) concentration in the aqueous solution. In both
cases the product of the concentrations of the potential-determining ions
in the solution (Ag^+ and I^-; H^+ and OH^-) is a constant. The electric-
potential difference corresponding to such a double layer (double-layer
potential) is:

$$\psi = \frac{kT}{ve} \ln \frac{c_+}{c^0_+} = -\frac{kT}{ve} \ln \frac{c_-}{c^0_-}$$

where c_+ and c_- are the concentrations of the potential-determining ions
in the solution, and c^0_+ and c^0_- are the concentrations of these ions for
the zero point of the charge. This zero point of the charge (or isoelectric
point) in the case of AgI/water is realized for (roughly) $c_+ = 10^{-6}$

eq/liter and $c_- = 10^{-10}$ eq/liter; in pure water $(c_+ = c_- = 10^{-8})$ the double-layer potential is therefore negative $(\psi = -116$ mv$)$. By raising the I^- concentration in the medium to $c_- = 10^{-5}$, the value, $\psi = 5 \times (-58) = -290$ mv, is reached. This concentration is roughly the free iodide concentration in a dialyzed AgI sol, obtained by preparing a negative sol by precipitation of $AgNO_3$ with an excess of KI, and subjecting it to electrodialysis.

If neutral electrolytes are added to such a system the double-layer potential is not altered, provided that no secondary changes occur which vary the concentration of the potential-determining ions. This condition is sufficiently fulfilled in diluted sols.

In the above, we have mentioned that the addition of electrolytes has the effect of increasing the capacitance of the double layer. Hence the condition where ψ is constant, independent of the concentration of neutral electrolyte, means that in a system such as a dialyzed AgI sol or an Al_2O_3 suspension, the double-layer *charge* is no longer a constant, but must increase with increasing electrolyte concentration. We see here a fundamental difference from a double layer of type (A), where the double-layer charge is roughly a constant determined by the amount of capillary-active ions in the interface and where the addition of electrolyte causes a decrease of ψ.

The situation may be different, however, in concentrated sols. This may be shown for the case of the dialyzed AgI-sol, which can be easily concentrated during the electrodialysis process by electrodecantation, and for which sol concentrations of 10 to 50 per cent AgI can readily be obtained. A dialyzed AgI-sol can be denoted by the symbol

$$\boxed{\text{AgI}} \;\; I^- \mid H^+$$

expressing the fact that the negative charge is present as an excess of I^- ions in the AgI surface and is balanced by a swarm of H^+ ions in the liquid surrounding the particles. In such a sol,[5] containing particles of about 50 mμ diameter and therefore comprising roughly 10^6 AgI "molecules," the charge of one particle is of the order of 500 elementary charges and the excess of I^- present in the double layer is consequently of the order of 0.05 per cent of the total amount of AgI. Hence we see that in a concentrated sol, containing for instance one mole of AgI (235 gm) per liter, the amount of iodide in the double layer amounts to 5×10^{-4} mole/liter, which is considerably more than the equilibrium concentration of free iodide in the sol medium $(\sim 10^{-5}$ mole/liter$)$. If neutral electrolyte is added to such a concentrated sol and the capacitance of the double layer is thereby increased, the amount of free iodide in the sol medium is by far insufficient to build up the addi-

tional double-layer charge necessary for the double-layer potential of the original sol. Contrarily, the effect of the neutral electrolyte will be that the free iodide concentration is lowered until a new equilibrium is established, in which the total double-layer charges are increased by a minor fraction corresponding to the increased adsorption of I⁻, and the double-layer potential is lowered to a considerable extent corresponding to the lower I⁻ concentration in the sol medium.[6] Hence, in these concentrated sols we again approach the conditions prevalent in systems where the double layer has been formed according to mechanism (A), and where the double-layer charge is virtually a constant independent of the concentration of neutral electrolyte.

Summary. Summarizing our considerations about the double layer of a single sol particle, we state that it is desirable to distinguish between a double layer due to the adsorption of capillary-active ions [type (A)] and one due to the distribution equilibrium of potential-determining ions [type (B)]. We have also investigated the behavior of double layers of both types under the influence of neutral or indifferent electrolytes, as an introduction to a treatment of the problem of stability and flocculation. It appeared that an increase of the electrolyte concentration, which has the effect of suppressing the thickness of the double layer, lowers the ratio ψ/σ in the two cases in a different way (σ = double-layer charge):

Type (A): σ constant, ψ decreases
Type (B): ψ constant, σ increases

Only in the case of a very concentrated sol of type (B), in which the total double-layer charge exceeds the amount of potential-determining ions in the sol medium, are the conditions holding for a sol of type (A) more or less restored.

Of course there may be cases where both mechanisms act together in the formation of the double layer, or even counteract each other. As an example of the latter type we may mention the influence of a positive capillary-active ion upon a negative AgI sol or As_2S_3 sol. In this case the effect of the "electrolyte" is clearly a decrease of the original double-layer charge. The same holds if a negative AgI sol is treated with a small amount of $AgNO_3$, precipitating the I⁻ excess in the system. Such electrolytes cannot be considered as examples of indifferent electrolytes. These cases can easily be treated according to the principles outlined above. In a study of the influence of electrolytes upon the electric double layer and the stability of the sol they can be considered as special cases. What follows is restricted to the more general problem of the effect of indifferent electrolytes. The latter are defined as electrolytes

producing ions which have neither an appreciable adsorbability nor the ability to act as potential-determining ions.

(II) Interaction of Sol Particles

We now pass to the problem of the interaction of sol particles. This is clearly a very fundamental problem in colloid chemistry, as the forces acting between particles determine what will happen when Brownian motion causes two particles to meet. If the particles attract each other, the result will be an agglomeration of the particles and ultimately a flocculation of the sol. If, on the other hand, repulsion forces are predominant, or at any rate predominant over a certain distance so that a potential barrier of sufficient height is present, the particles will separate again and therefore keep their individual motion. In the latter case we are dealing with a stable sol. We know from experiment that these forces can be influenced by small amounts of electrolytes in the sol medium. More especially, the markedly different action of ions of different valency, as laid down in the well-known Schulze-Hardy rule, is a striking feature of this effect.

A satisfactory explanation of these phenomena, in terms of attraction and repulsion forces between the particles, was only recently obtained. For this theoretical development two major steps in the theory of the interaction of sol particles had to be made. One was the theory of van der Waals-London forces; another was the theory of the interaction of two double layers.

De Boer [7] and Hamaker [8] have called the attention of colloid chemists to the importance of the London-van der Waals attraction forces between particles. They showed that the van der Waals-London attraction forces between atoms, as a consequence of their additive character, add up to an attraction between two colloid particles acting over an appreciable distance. Beginning with London's equations (for the attraction between two atoms), they calculated the order of magnitude of these forces, and more especially Hamaker gave equations for the interaction as a function of the particle distance for different shapes of the particles. The attraction force between two plate-shaped particles oriented parallel to each other, for instance, declines according to a d^{-3} law, as long as the particle distance d is small in comparison to the dimensions of the particles. Recently, Casimir and Polder [9] have improved the London theory and obtained equations showing a somewhat more rapid decay, especially for distances larger than 10^{-6} to 10^{-5} cm.

These van der Waals-London attraction forces are virtually independent of the composition of the sol medium as long as this consists of a dilute aqueous solution. Hence they cannot explain directly the influence

of electrolytes upon the properties of the sol. But it seems highly probable that these forces play an important part in the problem and, more particularly, are responsible for the fact that under certain circumstances an agglomeration of the particles occurs. We must assume that these attraction forces are always active, i.e., also in the case of a stable sol, where obviously their action is prevented or counteracted by forces which must clearly be of a repulsive nature.

About ten years ago a first approach to this problem was made by Frumkin [10] and Derjaguin,[11] and independently by Langmuir.[12] These authors investigated the problem of the interaction of the double layers, starting with the reasonable assumption that the nature of the forces preventing agglomeration under favorable conditions was to be found in the electric double layer of the particles. Levine [13] also contributed to the problem, though his first attempts in this direction failed.

Finally, Verwey and Overbeek [14] made an extensive study of the problem of the interaction of sol particles. They extended the theory of the interaction of two double layers and applied the results to colloids. These investigations were carried out during the war years and published in book form after the war. Following is a brief review of this work. The reader is referred to the book cited above [14] for a more detailed picture of the theory.

A semi-quantitative theory of the interaction of two interpenetrating double layers can be given only on the basis of a more or less satisfactory theory for a single double layer. Various possibilities as to such an underlying theory are available. A number of authors have tried to work with the theory of Debye and Hückel, which is an approximation of Gouy's theory for very small values of the double-layer potential and has been developed for solutions of electrolytes. It appears, however, that this theory is usually a very poor approximation for colloids, where the particle charges are generally so much higher than in the case of simple electrolytes. Stern's theory of the double layer gives a somewhat more detailed picture of the distribution of the charges and of the corresponding electric-potential function than does Gouy's theory. In the work of Verwey and Overbeek, the theories of both Gouy and Stern were utilized, although mainly Gouy's view because of its more simple mathematical treatment; the Debye-Hückel approximation was used only in special cases where this simplification seemed justified. In the following we will give only examples which have been based on Gouy's theory of the double layer, and which hold for the case of relatively large particles for which the thickness of the double layer is small in comparison to the particle dimensions. We will accordingly work with the simple concept of the flat double layer and neglect the influence of the finite dimensions of the ions and of specific forces between these

ions and the particles, which may become important when the ions come very close to the particle surface. These restrictions imply, for instance, that our considerations should not be extended to systems for which the absolute value of the double-layer potential is too high (say 200 mv or higher).

We will now consider what will happen to the double layers when two particles meet. Figure 1, curve $2d = \infty$, will show the electric potential in one double layer as a function of the distance from the particle; in a solution containing only small amounts of electrolytes (say 10^{-5} eq/liter) $1/\kappa$ will be of the order of 10^{-5} cm. In exactly the same way as for a single double-layer, we can easily derive the equations for the
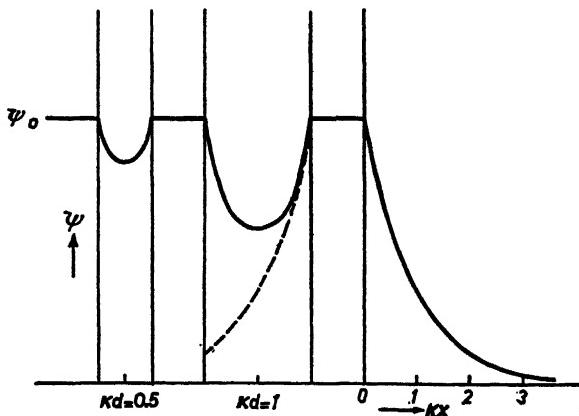
Figure 1. The electric-potential function for a single flat double layer (ψ_0 equals 51.2 mv) and for two interpenetrating double layers (1 — 1 valent electrolyte): κd equals 0.5, κd equals 1, and κd equals ∞.

electric-potential function of two double layers which belong to two parallel surfaces and are at such a short distance ($2d$) from each other that the diffuse charge layers overlap. Curves $\kappa d = \frac{1}{2}$ and $\kappa d = 1$ of Figure 1 are examples of this electric-potential function for two different values of the distance. One will notice that we have assumed that the double-layer potential (ψ for $x = 0$, or ψ_0) is independent of the particle distance. It follows from the equations for the charge distribution that in this case the double-layer charge is reduced by the interaction. Figure 2 gives σ, the charge of one double layer per cm², as a function of the particle distance, and shows that the double-layer charge drops sharply towards zero when the surfaces come close together.

The assumption that ψ_0 will be a constant independent of the particle distance, and that accordingly σ is reduced in the region between the particles when they are brought together, is certainly the most appro-

priate one for a double layer of type (B), as discussed in section (I). It corresponds to the assumption of thermodynamic equilibrium every-where in the double layer. It seems justified to assume that the distribution equilibrium of potential-determining ions is established so rapidly that it can easily follow the variations of particle distance due to the comparatively slow Brownian motion. Hence what actually occurs when two negatively-charged AgI particles in an AgI-sol meet and two crystal faces approach each other, is that a certain fraction of the excess of I⁻ ions responsible for the negative surface charge will go
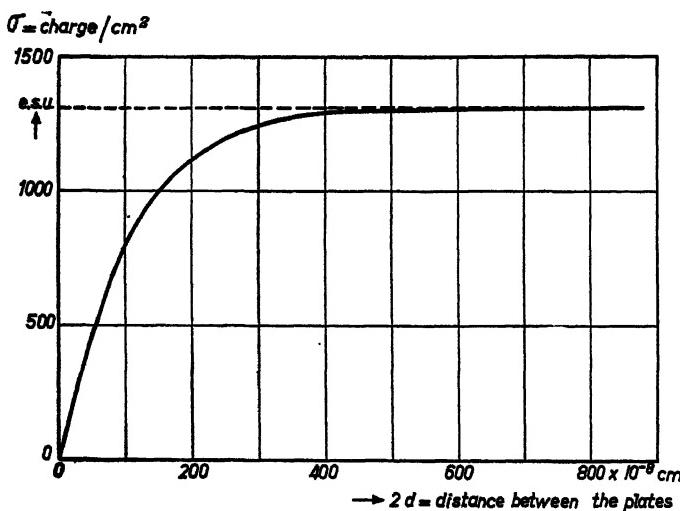
Figure 2. The charge of each double layer for a system of two interpenetrating flat double layers as a function of the distance (2d), for 1-1 valent electrolytes. ψ_0 equals 51.2 mv, $1/\kappa$ equals 1×10^{-6} cm.

from the surface into the solution and there neutralize part of the positive space charge squeezed between the two adjacent surfaces. This temporary adjustment of the charge as a consequence of the interpenetration of the diffuse charge layers is obviously due to the fact that in the new situation the double layers can no longer develop completely, merely because of lack of space. These counter ions need this space because of their thermal motion and their tendency to spread over the whole solution despite the electrostatic attraction by the surface charge. Hence the reduction of the charge is a direct result of the "osmotic" forces which are responsible for the extension of the diffuse charge layer of a single double layer.

For a double layer of type (A), one might expect that in a particle encounter the charge would be a constant independent of the distance of

the surfaces. In this case the double-layer potential would be raised by a decrease of the particle distance and would locally reach a value ∞ when the particles finally touch. Such an extreme condition would certainly not be very probable, and in many cases we may infer that also for double layers of this type the assumption, $\psi_0 =$ constant, is approached, as the adsorbed capillary-active ions carrying the surface charge will be more or less pushed away from the adjacent surfaces towards other parts of the surface of the particle or droplet. Hence our considerations will be based on the assumption $\psi_0 =$ constant, although other assumptions can easily be incorporated into the theory and do not lead to fundamentally different results.

The next step is the evaluation of the force between the particles, which is a result of the changes in the double layers when the diffuse charge layers interpenetrate. Or rather, the relative potential energy of the two particles will be considered, as this quantity can be more directly evaluated and is used directly in the potential curves considered in the following. These "potential curves," which have been used by Lennard-Jones, J. H. de Boer, and others in discussing theoretical problems, appear to be a convenient means of treating problems of this nature. The force can be read directly from these curves as the (negative) derivative with respect to the particle distance.

The potential energy is obtained directly if we know in what way the free energy of the system of two interpenetrating double layers is changed by a variation of the particle distance. As the interaction is effectively a partial suppression of the double layer, we can expect that the free energy of a system of two interpenetrating double layers is higher than that of the two separate double layers. When. a particle is immersed in the liquid the double layer is formed autogenously and the formation of the double layer is therefore accompanied by a decrease of the free energy. From this we can already infer qualitatively that the result of the interaction will be a repulsion. It can be shown that the free energy of a double-layer system can be split up into two parts, an electrical and a chemical part. The chemical part is the free energy gained in the formation of the double layer by the transition of the potential-determining ions from one phase to another, for instance the transition of the I⁻ ions in the case of AgI/water from the state of solution to the state in which it is incorporated into the crystal surface. This gain of free energy in the double-layer formation process is always larger than the loss of free energy associated with the build-up of the charges, which counteracts a further transition of I⁻ ions to the state of lower thermodynamic potential when the final equilibrium is reached. Inversely, chemical work must be supplied when the two particles are brought together and I⁻ ions have to be transferred to the

solution, and though a certain amount of electrical free energy is gained in the same process, the former will determine the sign of the total amount of work to be done. It will also be clear that this amount of work will generally be larger in the case of a double layer with a large capacitance than in one with a low capacitance (as present, for instance, in an emulsion which is not stabilized by an emulsifier).

Details of the derivation of an equation for free energy will not be given here. It is perhaps the most intricate part of the theory and at
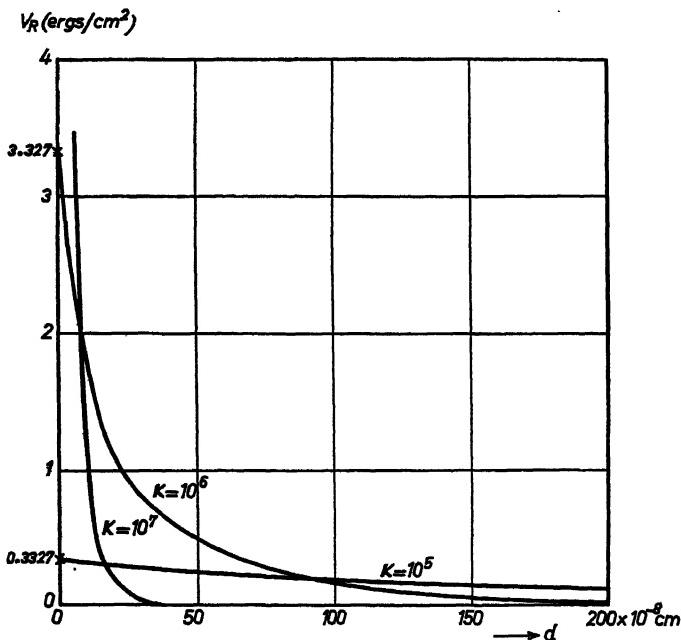
Figure 3. Repulsion potential for 1-1 valent electrolytes, assuming that ψ_0 equals 153.6 mv (z equals 6), for different concentrations.

any rate has been a serious obstacle to a number of authors who have been considering this problem. One way is to follow the method already used in the theory of Debye and Hückel in calculating the total amount of work in an imaginary process in which the charges of all ions of the system are reduced reversibly in infinitesimally small steps. It is also possible to make use of more general thermodynamic considerations (generalized Lippmann equation). Both methods lead ultimately to the same result (Casimir, Levine), but the former leads directly to a more convenient expression.

Figure 3 gives a few curves showing the repulsion energy derived in this way for two flat double layers, as present between the faces of two

parallel plates immersed in the electrolyte solution. This approximates the case of two colloid particles in a sol medium. The energy is expressed in ergs/cm²; the graph holds for 1-1 valent electrolytes, for $\psi_0 = 153.6$ mv (or $z = \dfrac{e\,\psi_0}{kT} = 6$), and shows the effect of different electrolyte concentrations: $\kappa = 10^7$ corresponds to a concentration of 10^{-1} eq/liter, $\kappa = 10^6$ to 10^{-3} eq/liter, etc. It is seen that the repulsion potential, V_R, for short distances increases with increasing electrolyte concentration. A much more important point is, however, that for larger electrolyte concentra-
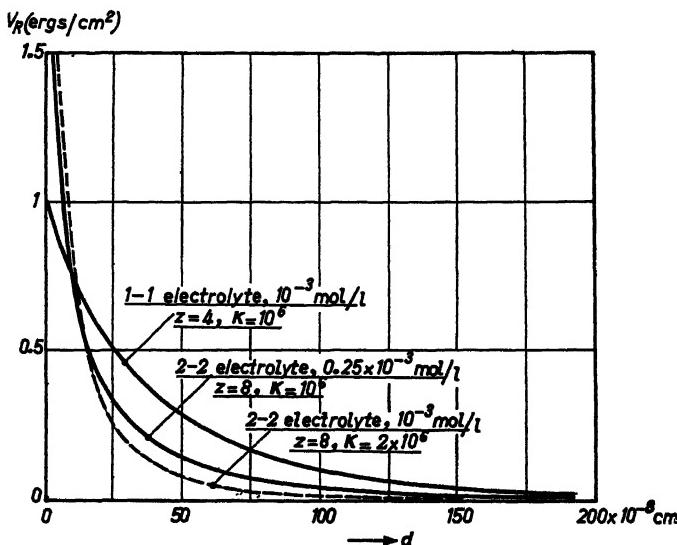
Figure 4. Repulsion potential as a function of the plate distance (2d) for comparable concentrations of a 1-1 and a 2-2 valent electrolyte. (ψ_0 equals 102.4 mv; z equals 4 and 8, respectively.)

tions the repulsion sets in at much smaller particle distances, which is obviously a consequence of the circumstance that in more concentrated solutions the double layers extend less far into the solution.

Figure 4 shows the effect of valency of the ions in solution and compares the $V_R(d)$ for 1-1 valent electrolyte and 2-2 valent electrolyte of comparable concentration (same κ or same molarity). The graph holds for $\psi_0 = 102.4$ mv, or $z = \dfrac{ve\,\psi_0}{kT}$, equal to 4 and 8 respectively. We observe that an increase of the ionic charge has the same effect upon the V_R curve as an increase of concentration.

Figure 5 contains a whole set of repulsion-potential curves on a logarithmic scale, for 1-1 valent electrolytes in the solution and for dif-

ferent values of the double-layer potential: from $\psi_0 = 25.6$ mv $(z=1)$ to $\psi_0 = 256$ mv $(z=10)$ upwards. By using a different scale for V_R, as indicated in the graph, the curves are applicable to different values of κ or of the ionic concentration. From these curves we see that V_R declines roughly according to an exponential law, more especially for small inter-
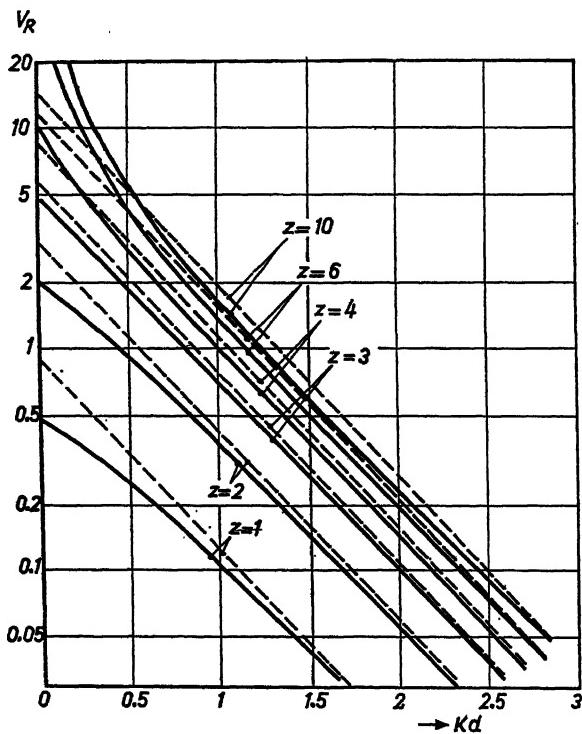
Figure 5. Repulsion potential on a logarithmic scale as a function of the distance for 1-1 valent electrolytes and different values of the double-layer potential (z equals 1 or ψ_0 equals 25.6 mv, etc.). Full curves are "exact" curves; dotted curves represent approximate equation (1).

action $(d > 2/\kappa$, or $\kappa d > 2)$. The dotted lines represent an approximate explicit expression for V_R, which can be derived for large values of d in comparison to $1/\kappa$ $(\kappa d \gg 1)$. This equation reads:

$$V_R = \frac{64 nkT}{\kappa} \cdot \gamma^2 \cdot e^{-2\kappa d} \tag{1}$$

where κ has the meaning mentioned earlier and

$$\gamma = \frac{e^{z/2} - 1}{e^{z/2} + 1}$$

This equation for V_R, apart from the factor γ^2, had already been derived by Frumkin and Gorodetzkaja.[10] It is a useful equation as it confirms the fact that V_R for large distances declines exponentially, and the more rapidly in proportion as $1/\kappa$, (the thickness of the double layer) is smaller. The equation will be used later on for the derivation of a simple flocculation law.

These repulsion-potential curves must now be combined with van der Waals-London attraction-potential curves in order to have a complete picture of the effect of the simultaneous action of both forces. A fundamental difference between these two types of forces is that the van der Waals-London attraction declines according to a negative power of the distance, whereas the repulsion declines according to a negative e-power. This means that for short distances the repulsion potential, starting with a finite value for $d=0$, is theoretically always smaller than the van der Waals attraction potential, starting with an infinitely large or at least a very large value for particle contact. At any rate, this relationship should hold for particles of a simple shape, such as spherical particles, or for plate-shaped particles oriented parallel to each other. Hence in these cases we should expect agglomeration of the particles once they have been brought sufficiently close together. If in stable sols the repulsion forces due to the double-layer interaction can prevent this agglomeration, we must expect that they must be able to create a potential barrier for the particles at some distance from each other. This is

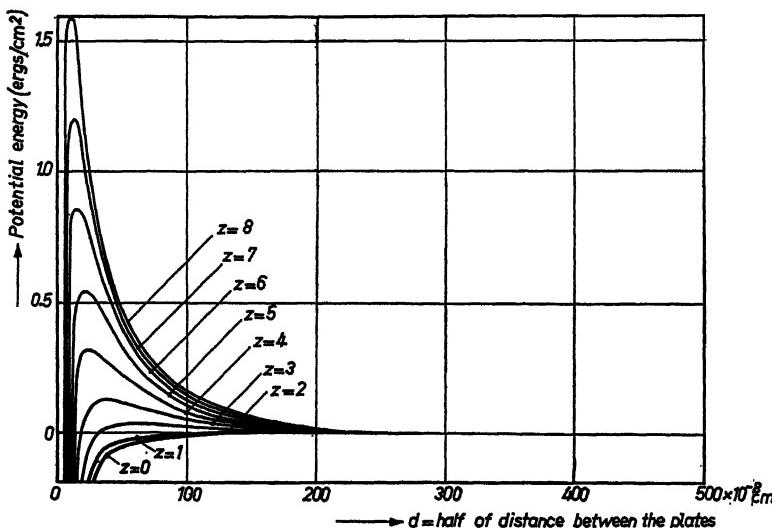Figure 6. Total-potential curves for A (van der Waals constant) $= 2 \times 10^{-12}$ and κ equals 10^6 (1-1 valent electrolytes.) ψ_0 varies from 25.6 mv to 8×25.6 equals 204.8 mv. The curve z equals 0 (no double layer) is also represented.

exactly what is obtained on the basis of the theory outlined here for reasonable values of the van der Waals constant and the double-layer potential. For a sufficiently low electrolyte concentration the repulsion potential is generally larger than the attraction potential over a certain distance region, and accordingly the total-potential curve shows a more or less pronounced maximum. This is illustrated in Figure 6, showing a set of curves for different values of the double-layer potential, for A (a factor determining the van der Waals-London constant)$=2\times10^{-12}$
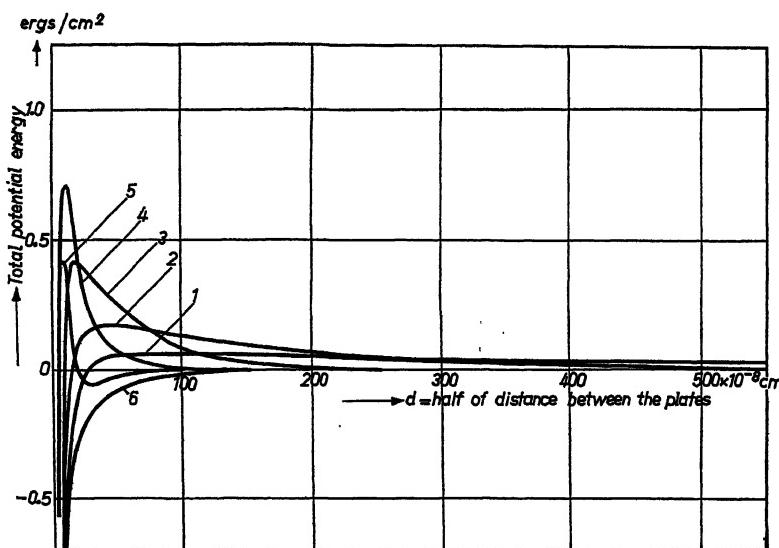
Figure 7. Total-potential curves, for A equals 10^{-12} and ψ_0 equals 102.4 mv (1-1 valent electrolyte), for varying values of the ionic concentration:

curve	1	2	3	4	5	6	
κ =	10^5	$10^{5.5}$	10^6	$10^{6.5}$	10^7	$10^{7.5}$	cm^{-1}
n =	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	eq/liter

and $\kappa=10^6$ (1-1 valent electrolyte). The figure also contains the curve $z=0$, representing the case of complete absence of a double layer, where we have only van der Waals attraction. The effect of electrolyte concentration, for a given value of ψ_0, is illustrated in Figure 7 ($A=10^{-12}$). It is seen that the very broad maximum, found for low electrolyte concentration, changes for higher concentrations into a more narrow though higher maximum at shorter particle distances; for concentrations above 10^{-2} eq/liter, it is lowered again, and finally disappears. In the last case the repulsion potential, though for short distances increasing with increasing concentration (Figure 3), is only active in a distance region where the van der Waals-London attraction predominates entirely.

A complete theory of the flocculation of colloids can now be given on the basis of these potential curves. Roughly speaking, we may say that a sol will be stable when the maximum is of the order of several times kT. For the detailed theory, references should be made to the book mentioned above. Here we will only give some general conclusions for the case where the particles are not too small. Then the simplified assumption may be made that a sol can be called stable when a particle distance can be found where $V_R > V_A$. For plate-shaped particles oriented parallel to each other, the van der Waals-London potential for the relevant distance region can be represented by a d^{-2} law. Combining this with the approximate equation (1) for the repulsion potential, we can easily derive for the transition between flocculating and stable sols the following criterion:

$$n = \frac{107\,\epsilon^3 k^5 T^5 \gamma^4}{A^2 (ve)^6}$$

Accordingly, we find for the flocculating concentration c (in millimoles per liter) of a v-v valent electrolyte at room temperature

$$c = 8.10^{-22}\,\frac{\gamma^4}{A^2 v^6} \tag{2}$$

A number of interesting conclusions can be drawn from this equation. In the first place it shows in a simple way that the flocculation con-
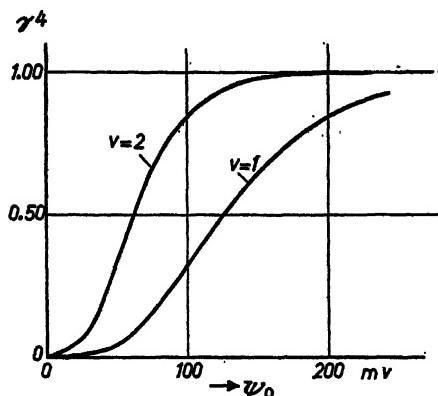
Fig. 8. The quantity γ^4, where γ equals $(e^z - 1)/(e^z + 1)$ and z equals $ve\,\psi_0/\kappa T$, as a function of ψ_0, for v equals 1 and v equals 2.

centration increases with increasing double-layer potential and therefore increasing value of γ. This concentration decreases, however, with increasing value of the van der Waals-London attraction.

It is interesting to study the influence of the double-layer potential in some detail. For that purpose Figure 8 and Table 1 show the quantity γ^4 as a function of ψ_0 for $v=1$ and $v=2$. It is seen that γ^4, and therefore

c, increase sharply with increasing ψ_0 as long as $v\psi_0$ has a low value. In this case we can easily derive an even simpler approximate equation relating ψ_0 and c at incipient flocculation, as for $\frac{z}{2} < 1$:

$$\gamma = \frac{1 + \frac{z}{2} + \cdots - 1}{1 + \frac{z}{2} + \cdots + 1} \sim \frac{\frac{z}{2}}{2} = \frac{z}{4} = \frac{1}{4} \frac{ve\psi_0}{kT}$$

TABLE 1

ψ_0 (mv)	γ^4	
	$v = 1$	$v = 2$
51.2	0.045	0.34
102.4	0.34	0.86
153.6	0.66	0.98
204.8	0.86	1.00

Hence we find instead of equation (2):

$$c = \frac{10^{-22}}{32A^2v^2} \left(\frac{e\psi_0}{kT}\right)^4$$

and we obtain the very simple result that the flocculating concentration is approximately proportional to the fourth power of the double-layer potential:

$$c = \text{constant} \cdot \frac{1}{v^2} \cdot \psi_0^4 \tag{2a}$$

Thus we have derived the theoretical foundation of an empirical relationship, formulated by Eilers and Korff,[17] and occasionally used in colloid chemical practice (e.g. Graham and Benning [18]):

$$\frac{\zeta^2}{\kappa} = \text{constant}$$

Here ζ is the so-called electrokinetic potential, as determined by electrophoresis. When the condition mentioned above ($v\psi_0$ is small) is fulfilled, this quantity ζ will not differ much from ψ_0. As κ^2 is proportional to cv^2, we arrive at the interesting result that the rule of Eilers and Korff is equivalent to our equation (2a), directly following from the theory for small values of $v\psi_0$.

For larger values of ψ_0, however, especially for valencies higher than 1, γ^4 approaches unity, and therefore the flocculation value is no longer very sensitive to the value of the double-layer potential. This leads to a very simple limiting law for larger values of ψ_0, relating the influence of the valency of the electrolyte to the flocculation concentration. Equa-

tion (2) contains the valency v both explicitly and, in the quantity γ, implicitly. If γ^4 approaches unity, however, it becomes more and more independent of v, and the concentration c is therefore merely proportional to v^{-6}. Hence we find in this case that the quantities of 1-1 valent, 2-2 valent and 3-3 valent electrolytes needed to flocculate a lyophobic sol or suspension are in a ratio

$$1 \; : \; (\tfrac{1}{2})^6 : (\tfrac{1}{3})^6 \quad \text{or} \quad 100 : 1.6 : 0.13$$

This result of the approximate theory stands in very good agreement with colloid chemical experience as formulated long ago in the well-known Schulze-Hardy rule.

In the above, it has tacitly been assumed that the double-layer potential ψ_0 does not change appreciably by a change of the electrolyte concentration or of the valency type of the electrolyte used. In section (I) we have seen that this is true only for sols or suspensions which have a double layer of type (B) built up by a distribution equilibrium of potential-determining electrolytes, and in the case of a sol the additional condition has to be fulfilled that it should not be too concentrated. In colloid systems where the double-layer *charge* is more or less a constant with varying electrolyte concentration, we must take into account that the double-layer potential is lowered by the addition of electrolytes. It appears, however, that the flocculating concentrations correspond roughly to concentrations of the same double-layer capacitance, so that the simple law derived above will still be valid, provided of course that the potential is not lowered to such an extent that γ^4 differs appreciably from unity.

In all cases where the double-layer potential is so low that the approximation mentioned above is not allowed, the concentration ratio between electrolytes of different valency at the flocculation limit will be less than given by the sixth-power rule. In such cases, the effect of the v^6 in the denominator is partly balanced by a larger γ for larger v in the numerator of equation (2).

Finally, we can derive from equation (2) an approximate value of A, while making use of experimental values for the flocculation concentration. For 1-1 valent electrolytes this concentration is of the order of 100 to 200 meq for most lyophobic sols. Assuming $\gamma = 1$ and $c = 200$, we find that

$$A = \sqrt{\frac{8 \times 10^{-22}}{200}} = 2.10^{-12}$$

This value is wholly within the limits set by the quantum mechanical theory of van der Waals-London forces.

Conclusions

These considerations may suffice to show that colloid chemistry is entering a new stage, where it appears possible to explain the behavior of colloids quantitatively with the aid of a number of fundamental physical phenomena such as the van der Waals-London attraction between atoms and the thermodynamics of phase boundaries. The next step in colloid research should be a more careful study of the electrical properties of the double layer for a well-defined object, as it is clear that the Gouy-Chapman picture, although very useful as a first approximation, is too simple. Stern's picture offers possibilities for an improved theory, but before this model can be used efficiently in calculations given above, more experimental data are needed. A first attempt in this direction can be seen in the work by Rutgers on the electrokinetic phenomena in glass capillaries.[15]

Another complication has recently been indicated by investigations on the system AgBr/water, showing that the charge on the AgBr particles is certainly not a surface charge but extends over a certain distance into the crystal.[16] This means that the excess of Br^- in a negative sol will be present as vacant Ag^+ lattice points in the AgBr lattice in the neighborhood of the surface of the particles. Such a charge will adjust much less rapidly when the conditions in the solution are changed, as this adjustment depends on the ionic diffusion velocities affecting the vacancies. We cannot consider here the consequences of this complication, since we do not know whether similar phenomena occur in other colloid systems.

REFERENCES

1. Verwey, E. J. W., *Trans. Faraday Soc.*, **36**, 192 (1940).
1a. Verwey, E. J. W., *Proc. Acad. Sci. Amsterdam*, **53**, 376 (1950).
2. Gouy, G., *J. phys. radium*, 9 (4), 457 (1910); Chapman, D., *Phil. Mag.*, 25 (6), 475 (1913).
3. Lange, E., and Berger, R., *Z. Elektrochem.*, **36**, 171 (1930).
4. Verwey, E. J. W., *Chem. Revs.*, **16**, 363 (1935).
5. ———, and Kruyt, H. R., *Z. physik. Chem.*, A167, 149 (1933).
6. de Bruyn, H., Thesis, Utrecht, 1938; Kruyt, H. R., and Overbeek, J. T. G., *Trans. Faraday Soc.*, **36**, 110 (1940).
7. de Boer, J. H., *Trans. Faraday Soc.*, **32**, 21 (1936).
8. Hamaker, H. C., *Rec. trav. chim.*, **55**, 1015 (1936); **56**, 3, 727 (1937).
9. Casimir, H. B. G., and Polder, D., *Phys. Rev.*, **73**, 360 (1948).
10. Frumkin, A., and Gorodetzkaja, A., *Acta Phisicochim. U.R.S.S.*, **9**, 327 (1938).
11. Derjaguin, B., *Trans. Faraday Soc.*, **36**, 203 (1940).
12. Langmuir, I., *J. Chem. Phys.*, **6**, 893 (1938).
13. Levine, S., *Trans. Faraday Soc.*, 42B, 102 (1946); **44**, 833 (1948).
14. Verwey, E. J. W., and Overbeek, J. T. G., "Theory of the Stability of Lyophobic Colloids," Amsterdam, 1948.
15. Rutgers, A. J., Private communication (to be published in *Advances in Colloid Science*, IV).
16. Grimley, T. B., and Mott, N. F., *Trans. Faraday Soc.*, **43** (1947).
17. Eilers, H., and Korff, J., *Trans. Faraday Soc.*, **36**, 229 (1940).
18. Graham, D. P., and Benning, A. F., *J. Phys. Colloid Chem.*, **53**, 846 (1949).

THE EFFECTIVE DEPTH OF THE
SURFACE ZONE OF A LIQUID

J. W. McBain and J. C. Henniker

Dept. of Chemistry, Stanford University, Stanford, Calif.

SINCE 1912 and the work of Hardy, it has been recognized that molecules have a definite shape, and that in liquids as in solids, they lie together in preferred orientations. These concepts have become especially familiar in the case of monomolecular layers of insoluble films on water. This idea has been so much emphasized that the necessary corollary, namely, that the orienting effect of such an oriented surface must extend to molecules in deeper layers, has been largely overlooked. The present review brings to the attention of the reader some of the evidence that the surface zone of most liquids is many molecules deep. It is not necessary to assume long-range forces, but rather to remember that short-range forces between adjacent molecules must successively orient these relative to each other and thus produce a surface zone appreciably different from the bulk liquid. A more detailed review of the relevant experimental evidence has been given elsewhere.[29]

The existence of long-range forces and even equilibria between charged surfaces in a conducting medium is now generally accepted.[43, 61] The Dutch school emphasized that van der Waals–London attraction may be felt at distances of some thousands of Ångström units between the surfaces because the potential energy between extended surfaces falls off with the inverse square of the distance of separation. Hence, the attractive force between surfaces is appreciable at distances where the attraction between individual molecules would be completely negligible.

Lennard-Jones and Dent have shown [42] that the direct range of action of a crystal surface, such as potassium chloride, is so appreciable that the density of an argon atmosphere at a distance of 10Å may be increased severalfold. A surface of this kind, or a completely-oriented outside monomolecular layer of dipolar molecules, must have an important effect upon the close-packed molecules of the liquid beneath them, where the

natural tendency of the liquid molecules towards orientation merely has
to be reinforced. Thus, by relayed action, thick oriented layers can be
built up. It is only increasing disorder caused by Brownian motion or
molecular vibration that gradually brings this effect to an end. In the
special case of liquid crystals, it is a familiar fact that the effect of the
surface may be relayed, even to a distance of some centimeters.

In applying the equations of hydrodynamics, such as Stokes' law,[27] or
those of electrokinetics to movement of a solid in contact with the liquid,
it has always been found necessary to assume a rigid or a semirigid
adherent layer of solvent clinging to the solid in addition to that already
allowed for in the hydrodynamic equations.

Concepts Prior to 1900

To the ancients, the surface of a quiet pool of water must have been
the smoothest thing imaginable. Copper plates could be polished to make
excellent mirrors, but none reflected with the faithfulness of the surface
of a liquid. There can have been little reason to think that such a
surface was any more complicated than the ideal formulated by the
Greek pioneers of geometry: the boundary of a body having area but not
depth. Indeed, what more perfect illustration could the mathematician
have given than the shining surface of water.

But towards the end of the eighteenth century, the phenomena of
capillarity were being studied, and the molecular theory was gaining
ground. It was no longer possible to regard a liquid-gas interface as
geometrically smooth. The molecules were, and still are, supposed to have
a cohesive attraction for one another. This includes an attraction be-
tween neighboring molecules in the plane of the surface, and leaves an
unbalanced force tending to pull all molecules inward. This causes
bodies of liquids to assume a shape of minimum area, just as if contained
within a contractile skin.

At a time when so little was known about the forces between small
particles, there was no reason to deny long-range forces between mole-
cules at some distance from each other, and thus between molecules
exposed on the surface and those far below. For instance, in 1869,
Quincke [52] reported that the molecular forces responsible for capillarity
were operative as much as 800Å from a glass wall which was separated
from the liquid by a collodion film of this thickness. His technique was
to observe the contact angle of liquid drops on glass coated with films
of silver, collodion, etc. The contact angle remained the same as on clean
glass until a certain critical film thickness was reached; this thickness
being taken as the range of the attractive molecular forces originating
at the glass. It is noteworthy that until the kinetic theory of gases was

established, gas molecules were thought to repel one another, even across the relatively large intermolecular distances corresponding to the lowest pressures measurable.

Benjamin Franklin had measured the maximum area, and hence the minimum thickness, of a film of oil on a pond. In 1890 Rayleigh used the decrease in surface tension, as indicated by the stopping of the motion of camphor on water, to demonstrate that a film of olive oil on water was 16Å thick.[53]

The Period 1900 to 1920

Rayleigh's observation led the way to the study of surface films of insoluble substances on liquids. Langmuir's studies [40] centered upon the idea that the surface properties of liquids are actually the properties of the surface atoms, i.e., a matter of the outer Ångström or so. Thus, the concepts on the depth of influence of the surface of a liquid had returned almost to that of the ancient geometers, although in a highly sophisticated form. It was now felt that no forces originating at the surface could have an effect on a liquid at more than atomic distances, certainly not distances such as the length of a molecule. A liquid was a liquid, to the last atom of its being.

W. B. Hardy: Adhesion and Friction

The first modern investigator to break away from the monomolecular theory of liquid surfaces was Hardy. Anticipating later thinking on intermolecular forces, he postulated in 1912 a relayed effect passing from molecule to molecule "so deeply as to modify the molecular state of a skin some hundreds of microns in depth." [20] Subsequent data of many other workers place this estimate as several times too high, but qualitatively correct.

Fourteen years later this conjecture was supported by some surprising evidence from a study of adhesives. Not only did low-viscosity liquids exhibit a tensile strength, but the nature of the solid faces affected the adhesive layer. Hardy and Nottage,[23] using butyl, octyl, or undecyl alcohol between polished surfaces of steel and copper, found the adhesion between dissimilar metals to be the mean of that between the respective pairs of similar metals; not the minimum, as would be expected if the joint strength were determined at the metal surface.

Hardy also stated in the above paper that the same final maximum value of adhesion was attained whether the adhesive layer was formed by squeezing out an excess of liquid or by allowing the liquid to run in between dry surfaces in contact. The thickness of the liquid layer was

"a fraction of a millimeter." Significantly, the equilibrium thickness was reached far more quickly in increasing the liquid layer than in decreasing it. In the words of Hardy, "When the cylinder rises, fluid of low viscosity is drawing in, when it falls, it presses out lubricant whose molecules are locked in place by the attraction fields of the solids. It is the difference between drawing in a light spirit and expressing a jelly."

Subsequent work by Hardy and others has been with solid and plastic adhesives rather than simple liquids. The data are therefore not entirely à propos. But the adhesives were not crystalline solids, in which the effect of the wall might have been transmitted much as the growth of a single crystal is oriented. They were amorphous materials in which it is perhaps more reasonable to assume that the molecules were oriented by the wall while still in the liquid state.

The effect of the wall was demonstrated in a number of ways by McBain and his colleagues. Comparing the tensile strength of films of the adhesive alone with the strength of joints, it was found that the metal increased the strength of the adhesive severalfold.[45] In the same paper it was reported that the strength of a joint increased with diminished thickness. For example, joints of soft shellac between nickel surfaces were increased in strength seven times in diminishing the thickness from 0.010 to 0.003 inch. This is thought to be caused by the elimination of 0.007 inch of the relatively unorganized central part of the adhesive layer.

Studies with a number of metals showed [44, 45] that for a given adhesive the identity, and to a lesser extent the surface condition, of the metal faces influenced very strongly the strength of the joints. It was found that the strength increased greatly with the hardness, elasticity, and tensile strength of the metal and decreased with its atomic volume and compressibility.

One of the most striking demonstrations was made by Morrison and reported by Fuller.[19] A beeswax joint between two tungsten surfaces had an average tensile strength of 260 pounds per square inch; that between copper surfaces, 795 pounds per square inch. A similar joint between a copper and tungsten surface would be expected to show the strength of the weaker joint, were the strength determined solely at the surface. However, the actual strength was 600 pounds per square inch, showing that the copper had transmitted its influence across the adhesive layer. A similar result was obtained using de Khotinsky (a shellac creosote) cement between copper and tungsten.

In these investigations the thickness of the adhesive layer was not measured, but after breaking the joints, layers of visible thickness remained on each face. This would demonstrate the effective range of the forces in these adhesives to be of the order of microns.

Hardy and Nottage [24] also investigated joints between dissimilar metals, with comparable results. Copper and steel joined by solid paraffins and fatty acids in layers four microns thick showed a strength that was the mean, not the minimum, of the joints between similar metals.

Hardy's other major contribution to the study of long-range forces in liquids was his work on lubrication. He distinguished boundary lubrication, involving only a few layers of molecules between the sliding faces, and hydrodynamic lubrication in which the laws of fluid flow apply. He demonstrated static friction in what he regarded as thick layers of lubricant, itself evidence of a yield value in the liquid, and consequently of considerable intermolecular forces.

The actual depth of the liquid layer was not measured, but the indication is that it was significantly thicker than in boundary lubrication. In Hardy's experiments [21] the static friction between highly-polished glass surfaces, one plane and the other spherical, lubricated with long-chain fatty acids or alcohols, eventually reached the same steady value no matter how the system was assembled; this is considered to be the condition of boundary lubrication involving monolayers only. But the initial friction of the system was variable. Immediately after bringing the surfaces and fluid together, the friction coefficient was high (0.57 for caprylic acid, which will serve as a representative lubricant). Hardy considered that the liquid had not had time to orient and had been rapidly squeezed out, leaving only a poorly-oriented primary film. When, however, the glass surfaces were allowed to come to equilibrium with the liquid before being placed together, the friction was low (0.26), rising rapidly to 0.44. This period was interpreted as temporary hydrodynamic lubrication ending in boundary lubrication by a somewhat incomplete primary film. The friction subsequently dropped to the steady value (0.34), which corresponds to the complete monolayer. The initial low value (hydrodynamic) was absent between surfaces in equilibrium with only the vapor of the lubricant, which indicated poorly-oriented primary films in this case also.

In some of Hardy's later work on lubrication, the static friction coefficient was measured between a plane slider and a plane face. Under these conditions the normal pressure was low enough to insure viscous, rather than boundary, lubrication. Using several organic liquids, Hardy and Bircumshaw [22] found that after a "latent period" the friction coefficient reached a steady value that was in inverse proportion to the normal pressure on the slider. (This leads to the interesting conclusion that the tangential force required to move the slider was independent of load.) At higher pressure the friction coefficient was constant. Hardy considered this final condition to be boundary lubrication, dependent on a small number of molecular layers.

In a later paper, Hardy [25] measured the separation between the plane surfaces and found a gradual change in the friction coefficient for alcohol between steel surfaces as the film decreased from 70,000 to 4000Å. Bowden,[5] however, using alcohol between steel surfaces, found that the friction coefficient was negligible until a film thickness of 3600Å was reached, below which point it rose rapidly. He considered that the friction in the thicker films previously reported must have been caused by contamination of the liquid. Deryagin [9] pointed out that these data do not contradict his own evidence of deep surface orientation extending over 1000Å into a liquid. Later Deryagin and Smolinanskii [16] reported that the friction coefficient of films of paraffin oil containing aliphatic acids or alcohols could be used as a measure of adsorption. Films from 1000Å to 10 microns thick were involved in this study.

A wealth of evidence for mechanical strength in liquid films has accumulated since the pioneer work of Hardy. Unfortunately almost all of it is subject to the same uncertainty that has attached to Hardy's work, namely, that the solid surfaces were not ideally smooth, or that minute particles of dirt were responsible for the behavior attributed to multimolecular films. However, several investigations have been made in which the film of liquid was confined between a solid surface and a gas phase. In these cases, asperities or dust particles cannot decisively affect the results, which are, therefore, more significant and convincing.

In 1944 Deryagin and his colleagues [14] announced a method for comparing the viscosity of an oil near the solid surface with that farther away from it. A shearing force on a layer of oil was produced by blowing a stream of air over it in a narrow rectangular slit or trough. The thickness of the resulting film was estimated from interference patterns set up by the reflection of monochromatic light. Several oils, including a "Vaseline" oil containing 0.4 per cent oleic acid, did not show any anomaly in viscosity beyond 1000Å from the wall, but a turbine oil containing additions of aluminum naphthenate up to 2 per cent showed anomalous interference bands corresponding to a tenfold increase in viscosity within 5000Å from the wall. The same additive caused deviations from Newtonian viscosity in the bulk liquid. The authors interpreted the wall effect as an intensified or preferential development of a colloid structure in the proximity of the oil-metal surface.

The blow-off technique was improved [15] by using a narrow, wedge-shaped trough instead of a rectangular one. This produced an interference pattern which, when photographed, gave a direct representation of the viscosity of the oil as a function of the distance from the wall. The sensitivity of the method was later improved by blowing the air radially over the oil layer.[39]

Multimolecular Adsorption

Also among the first evidence of forces extending through a number of liquid molecules were cases of adsorption of vapors in greater quantity than that required to form a monomolecular layer. McHaffie and Lenher [47] in 1925 assembled a number of cases of multimolecular adsorption of water vapor on glass. They themselves obtained adsorption corresponding to from 1 to 184 monolayers of water, assuming the visible surface of the glass to be the effective surface for adsorption. Further data by Lenher in 1927 [41] showed water films 45Å thick on silica and 5300Å thick on glass; benzene films of 930Å on platinum and 110Å on glass. These films were all thermodynamically stable, since equilibrium was approached from both sides. It must be emphasized that the experiments were conducted at temperatures between the normal melting and boiling points of the adsorbed substance. This indicates that the adsorbed molecules are not as loosely bound to one another as in the bulk liquid.

Those who have not wished to accept the implications of multimolecular adsorption have raised several objections. One is that the water is held in the silicate lattice of the glass as it is in the thin membranes of the "glass electrode," where swelling of the glass can be demonstrated. This is improbable in the light of the equilibrium conditions cited above. Moreover, a swelling of platinum by benzene is hardly credible. Another objection is that the solid surface was not smooth, but was indented with submicroscopic fissures. This point has been examined by Joris and Taylor,[35] who have used a "roughness factor" of 4.75 for glass, obtained from radioactivity measurements with tritium, and have calculated Lenher's films to be 18 to 62 molecular layers. Another point that has been raised is that multimolecular adsorption may be irregular, concentrated at active centers on the solid. If this is true, the range of forces exerted by the solid through the adsorbate is even greater than that calculated on the basis of uniform multilayers. Multimolecular adsorption serves to show that an attractive force may make itself felt from molecule to molecule of a substance that in its bulk form is a liquid at the temperature of the experiment.

A most interesting and surprising example of multimolecular adsorption is that of helium vapor on glass at 1.72°K. Here there are no less than thirty layers of sorbed helium, while the vapor pressure is still only 90 per cent of saturation. This is ten times the number of layers observed on the same glass with neon vapor.[38]

Adsorption in superimposed layers is the basis of the well known isotherm that Brunauer, Emmet and Teller [8] announced in 1938. The B.E.T. isotherm has been found to fit the adsorption of many vapors

on finely-divided solids, which shows that multimolecular adsorption probably occurs also in the cases where no direct measure can be made of the area of the adsorbent.

Oriented Molecules

It is well-established that certain liquids have the bulk property of being oriented as "liquid crystals." The orientation readily observed by the extinction of polarized light, may be seen to proceed from the wall of the container until the whole mass of the liquid is doubly refracting. The influence of the solid wall was elegantly demonstrated by Zocher and Coper in 1928.[64] They brought about an orientation of the surface of a glass plate merely by rubbing it with filter paper or cotton. This can be communicated to a material placed upon the surface. (The orientation could be detected by evaporating on the glass an alcoholic solution of dye, the deposit of which was seen to be doubly refracting, whereas, on any untreated surface it was isotropic.) The induced orientation of liquid crystals was demonstrated by melting p-azoxyanisole on such an oriented glass surface, warming until the liquid became isotropic, and allowing it to cool; the double refraction in the liquid state was then found to follow the direction of the orientation of the glass plate.

Five years later Taylor and King [60] demonstrated double refraction in long-chain fatty acids as much as 5.7° above the melting point; whereas, the ordinary liquid is isotropic. They used the critical-angle method to measure refractive indices in relatively thick films of the liquid. The actual thickness of the doubly refracting part of the liquid was not measured, but it must have been at least one wavelength of light, or about 200 molecules, or several thousand Ångströms deep. The effect ceased abruptly at a certain temperature, which correlates with the behavior of lubricating oils reported by Brummage.[7]

Brummage obtained electron diffraction patterns from films of straight-chain organic compounds on metal faces, and studied the effect of temperature. He discovered a disorientation temperature, often well above the melting point, up to which the film gave a definite diffraction pattern. The disorientation temperature increased with film thickness up to about 20 molecular layers and then remained constant up to 100 molecular layers. This indicates that the metal surface had an orienting effect on the liquid film to a depth of thousands of Ångströms. The ability of the metal surface to orient molecules in the liquid phase was further illustrated by the fact that disorientation temperatures of films deposited from solution in isohexane were the same as those of films deposited by the Blodgett technique.[4] The latter are known to be as highly oriented as a crystal.[8a]

It seems reasonable to suppose that this orienting ability, possessed to an extreme degree by a few liquids, is to some extent the normal property of all liquids; that the influence relayed through p-azoxyanisole to a distance of centimeters differs only quantitatively from the influence relayed through tens or hundreds of Ångströms in common liquids, especially those whose molecules are much longer in one direction than in another. Indeed, x-ray examination has shown that in such liquids, as for example, hydrocarbons or alcohols, neighboring molecules lie side by side and end to end (cybotaxis), the disorder increasing rapidly with distance.[62] These liquids are already close-packed and the well-oriented monomolecular layer must greatly increase the cybotaxis to a considerable depth.

Blanketing Layers

There are some remarkable instances in which some such effect as has been described above is detected after passing through a layer of a material that has no orienting influence of its own. In one case the effect is directly observable as a change in the habit of growing crystals. In the other, the effect is biochemical and is not recognized as being determined by orientation.

The long-range influence on the growth of crystals was described by Bradley in 1937.[6] Ammonium iodide normally crystallizes as cubes. On the surface of mica it can be made to form tetrahedral crystals that are oriented in accordance with the crystal structure of the mica. When a film of cellulose acetate or rubber 1000Å thick was first deposited on the mica and the ammonium iodide was deposited on top of this, it again crystallized in the tetrahedral form. This indicates a propagation of the orienting effect of the mica through the organic film. The particular importance of this work lies in one experiment in which Bradley spread octane on mica in a film thick enough to show interference colors, and showed that the orienting effect was transmitted through this liquid layer. The possibility that the effect occurred via holes in the blanket layer was minimized by three observations: no cubic crystals were formed along with tetrahedrons; crystals of intermediate form grew on films too thick to allow formation of regular tetrahedrons. A film of gold 1100Å thick, but seen under the microscope to be porous, entirely inhibited tetrahedral crystals.

The second case of this kind is the extraordinary discovery of Rothen that biochemical reactions could be affected through hundreds of Ångströms of a blanket layer. The demonstration involving an immunological action was made as follows.[55] Films of antigen were plated on to a solid surface. These films are capable of adsorbing the homologous

antibody. The antigen was then covered with a blanket layer of an organic substance such as barium stearate, octadecylamine, or a plastic as much as 200Å thick. These substances alone will not adsorb the antibody. When the antibody was placed in contact with the blanket layer, it was found that the antigen was able to adsorb the antibody through the intervening layer. The possibility that this effect was caused by holes in the blanket layer was discounted, since the thickness of the blankets required for this effect depended on the nature and the thickness of the underlying layer of antigen.

Rothen [54] demonstrated the same kind of effect with an enzyme. In this case antigen bovine albumin was deposited on a metal plate and covered with an organic blanket layer. Thick blanket layers protected the antigen from the action of trypsin, but the enzyme was able to deactivate the antigen through thinner (65 to 150Å) layers. In this case also, the thickness of the blanket layer required for protection was a function of the number of layers of antigen beneath.

Rothen's effect involves a solid rather than a liquid film, and it seems that whatever orientation is responsible is more likely to have originated while the blanket film was liquid than after it had solidified. The entire series of experiments has been criticized on the basis that the blanket layers were porous. For instance Karush and Siegel [36] have emphasized the possibility that the underlying protein layer would be expected to be uneven and peaks on its surface would protrude through the blanket layer. It has been pointed out, however, by Iball [34] that monolayers of barium stearate, deposited by the Blodgett method, will bridge across the holes even of a wire gauze. Rothen's blanket layers which were deposited by the same method would thus bridge across any peaks in the underlying layer. Moreover, most of the protein would then be farther from the antibody than Rothen had stated, and the range of the effect perhaps greater.

Liquid as a Structural Element in Thin Layers

In the data cited under previous headings, the effect of the phase-boundary of a liquid has appeared either as an anchoring or as an ordering of molecules through many molecular layers of liquid. This orienting influence, transmitted from layer to layer of molecules, is strong enough in the following examples to prevent complete disruption by thermal agitation or by light mechanical pressure. In most of the examples now to be given, the liquid is water.

Most of the data have appeared in the last ten years; much of it depends on the interpretation of x-ray diffraction photographs. The work of Hofmann and Bilke on the structure of clays appeared in 1936.[32] They

found that as montmorillonite became hydrated, the repeating structure responsible for the x-ray diffraction pattern increased in thickness. This swelling, apparently caused by water wedging in between clay crystals, measured 10Å for calcium and hydrogen montmorillonite, and 20Å for the sodium clay. The relative thickness of the clay-unit crystal and the separating layer of water have been confirmed by electron microscopy by Shaw.[57] Various organic liquids, such as glycerol, may take the place of the water layer. In montmorillonite the layers were from 1 to 3 molecules thick, depending on the organic liquid adsorbed.[46]

Hardy's early experiments on friction had demonstrated that solid faces remained separated when covered with lubricant, but critics asserted that asperities, not the liquid, were responsible. Two independent groups justified Hardy in 1933.[2, 59] In one case a film of paraffin 100Å thick, and in the other case films from 300 to 1000Å thick, resisted the pressure of a surface of mercury, where the problems of asperities and foreign particles do not arise.

Equally positive results were obtained by Deryagin and his colleagues in 1939,[11, 12] in the film of liquid between a gas bubble and a sheet of mica. The film of water, which was measured from photographs of Newton's rings, thinned to 300Å in a few minutes, but did not diminish further in 24 hours. Moreover, by heating and cooling the system, the 300Å film was shown to be in equilibrium. Schofield[56] has evolved a theoretical justification for these results.

The validity of these results was challenged by Elton[17, 18] on both theoretical and experimental grounds. He attributed the long duration of the film to the effect of normal viscosity and electro-viscosity (the retarding action of a streaming potential built up by the presence of an electrical double layer on the wall of the channels). Elton presented experimental plots of the film thickness as a function of time for bubbles in benzene, ether, and $1N$ KCl (liquids having electrical double layers of negligible thickness), and showed that the plots could be extrapolated to a thickness of zero. His estimated experimental error was 50Å, so that his finding is not in disagreement with the statement of Deryagin and Kusakov that the film of alcohol on glass was too thin to be measured by the optical method in use. Elton showed no such plot for water, the liquid which both he and the Russian investigators had found to give the thickest films. The actual thicknesses were comparable: 180Å after 3 hours for water on quartz under 2800 dynes/cm^2 by the former; 300Å after periods from a few minutes to 24 hours for water on mica under 1500 dynes/cm^2 by the latter.

Elton, in his theoretical treatment of the problem, tabulated the effect of electroviscosity on the rate of approach of two plane solid surfaces in a liquid. He showed that, for instance, the time required for two plates

to come within 500Å is 20 times as great for an ionic as for a nonionic liquid. But this does not seem to be sufficient by itself to account for the experimental result that the thickness of a film of water after 24 hours was greater than that of a film of "Vaseline" oil after one hour, since the normal viscosity of the oil was probably many times that of water.

The first x-ray diffraction patterns of soaps were reported by Hess and his collaborators in 1937.[30, 31, 37] Since that time many colloidal electrolytes have been examined by the x-ray technique. It is now generally accepted that the patterns derive from plate-like micelles which consist of layers of long-chain ions with their nonpolar ends packed side by side. The structure may be likened to two pieces of pile carpet stacked with the pile sides together. Whether or not the micelles contain more than a single pair of layers is still open to question. It is possible that the x-ray interference is caused by arrays of micelles that have only transitory existence, like the cybotactic groups postulated by Stewart[58] to account for the diffuse x-ray patterns from pure liquids. These arrays are thought to be built up of pairs of layers with their polar surfaces facing one another. Whether or not these arrays are permanent, there is no doubt that the repeating structure responsible for x-ray diffraction contains water, presumably in the form of a layer between polar faces. The thickness of the water layer, and hence of the x-ray spacing, varies inversely as the soap concentration. For example, a water layer 42Å thick separates the soap layers in a 9.2 per cent solution of sodium oleate, the lowest concentration at which x-ray patterns were obtained.

Hughes, Sawyer, and Vinograd[33] demonstrated a still thicker water layer in micelles of potassium laurate. The longer of two x-ray spacings varied from 50 to 125Å, depending on dilution. The double length of the soap molecules could account for only 40Å, leaving 10 to 85Å due to water. When toluene was solubilized in the soap solution, the long spacings were further increased, indicating that the hydrocarbon had formed layers between the nonpolar faces of the soap layers, the same amount in each successive layer. Similar behavior was reported by Harkins, Mattoon, and Corrin in 1946.[26] They found that a mixture of isoprene and styrene increased the long spacing in a soap solution by 16Å. Recent experiments by Philippoff[51] have provided a striking confirmation of the above evidence by reversing the role of the aqueous and hydrocarbon components. A long spacing of 40Å was found in a 25 per cent solution of Aerosol OT (dioctyl ester of sodium sulfosuccinic acid) in decane. The Aerosol micelles can solubilize water to the extent of 220 per cent of the weight of the detergent. Moreover, the long spacing is then 170Å, which indicates an oriented water layer 130Å thick.

The same kind of structure has been demonstrated in a very different system by the x-ray work of Palmer.[48, 49] An emulsion of mixed nerve

lipids containing cephalin appeared to consist of bimolecular leaflets. The long spacing increased with decreasing concentration of the emulsion, and in a 25 per cent emulsion it reached 150Å. The water layer must have occupied at least 85Å, since only 65Å could have been accounted for by the lipid molecules. Ions had a profound effect on the structure, 0.076N calcium chloride being enough to expel the water layer entirely.

Palmer and his colleagues have produced new evidence of structure in intracrystalline water in their work on pectinic and pectic acids.[50] The water contents of the solid materials increased with the relative humidity of the environment, and concurrently the x-ray spacing between the polymer chains increased, though only from 6.1 to 7.4Å. Moreover, the crystallinity revealed by the x-ray patterns was perfected as the water content and the spacing increased. This suggests that the structure of the solid as a whole had been enhanced by intermolecular forces in the water. The behavior of the most methylated samples (10.9 per cent methyl ester content) was peculiar; the x-ray spacing remained nearly constant at all water contents. It was suggested that the prevalence of the hydrophobic methyl groups had prevented the chains from coalescing when water was removed. However, it seems to one of the authors * that this may have been caused by a more rigid water structure than in the less methylated material, since in the latter, the spacing at the highest water content exceeded the constant spacing of the high-methoxy acids. This is supported by the fact that the low-methoxy acids are not as water-soluble as the high-methoxy acids.

In an exceptionally careful and elaborate x-ray study of methemoglobin, Boyes-Watson, Davidson and Perutz [5a] showed that the layers of protein molecules making up crystals of the methemoglobin are separated by layers consisting of varying amounts of semirigid water. The solid, though not hard, was crystalline, and could only have retained its shape if a layer of water 15 to 25Å thick constituted part of the crystal structure. The water layer must not, however, be considered solid in the sense that ice is solid, since ions diffused readily into it and at the isoelectric point did not alter the layer spacing. On the other hand, if the water did not form a series of identical thick semirigid layers, the x-ray pattern and the crystallinity would not be possible.

A comparable investigation by Bernal and Carlisle [3] has revealed layers of water 78Å thick in the crystal structure of turnip yellow mosaic virus.

Wedging

Deryagin was the first to show that liquid molecules can wedge themselves between surfaces and separate them to multimolecular distances.

* J. C. Henniker

His term for the mechanism has been generally translated as "disjoining action," but the word is more literally "wedging apart." The concept was used in discussing the compressive strength of films of liquids mentioned previously, and in the development of a theory of the stability of weakly-charged colloids.[10, 13]

Beeck and his co-workers [1] also applied the idea of wedging of molecules, though their experimental basis was confined to lubrication studies. They discovered a sudden decrease in friction coefficient of a lubricated bearing above a certain speed. The new phenomenon was called wedging since it occurred only in the presence of excess oil and appeared to be caused by drawing-in of oil. The onset of wedging was also detected by a sudden increase in electrical resistance between the metal surfaces. It was correlated with molecular orientation as observed by electron diffraction. White oil, which is nonpolar, and ricinoleic acid, which adsorbs flat-wise, showed no wedging. Film thicknesses were not measured directly, but it seemed certain that wedging consists of the building up of thick oriented films.

The wedging postulated by Beeck differs from Deryagin's "wedging apart" in depending on rapid shear between layers of lubricant. However, in so far as the two processes are similar, one may well inquire if the peculiar properties of the oriented layers of lubricant are found to some extent in the surface zone of all liquids.

The outstanding peculiarity of layers of lubricants in bearings is their apparent fluidity in the direction of mechanical motion and rigidity in both directions perpendicular to it. This was clearly shown by Neale [47a] in 1943. It is as if the lubricant were organized as distinct laminae, like flakes of graphite, that slide over one another to build up multiple layers.

There are two other examples of a different but related effect, each an isolated observation, that would repay further study. While investigating the flow of a hydrocarbon oil in a narrow glass capillary, the present authors observed a 4 per cent drop in viscosity when a little lauryl alcohol was added.[28] The conditions were such that an oleophobic monolayer would be formed on the glass, as described by Zisman and his colleagues.[63] The mechanism postulated by Zisman is a close-packed monolayer adhering to the solid, having the methyl end groups toward the liquid.

The other case in which a monolayer has been shown to decrease the friction between a liquid and a solid was reported by Deryagin and Krylov [13] in 1944. They measured the viscosity of water in pores about 1000Å in diameter, and found that coating the pores with oleic acid resulted in a sevenfold increase in flow, which is ascribed to dehydration by the oleic acid. An aspect of this which occurred to one of the writers * is that a monolayer of oleic acid might entirely replace the electric double

* J. C. Henniker

layer, leaving no unbalanced charge to extend its influence out from the surface. The water in the capillary would be electrically neutral and would be bounded not by a negatively charged solid surface, but by a neutral surface consisting of the ends of hydrocarbon chains. The decrease in viscosity then is strongly suggestive of the phenomenon of boundary lubrication. It is noteworthy that in both these cases the coated solid was not wetted by the flowing liquid.

REFERENCES

1. Beeck, O., Givens, J. W., and Smith, A. E., *Proc. Roy. Soc. London,* 177A, 90-102 (1941).
2. Berdennikov, V. D., Bresler, S. E., Zelmanov, I. L., and Straut, E. A., *Physik. Z. Sowjetunion,* 4, 873-84 (1933).
3. Bernal, J. D., and Carlisle, C. H., *Nature,* 162, 139-40 (1948).
4. Blodgett, K. B., *J. Am. Chem. Soc.,* 57, 1007-22 (1935).
5. Bowden, F. P., *Physik Z. Sowjetunion,* 4, 185-96 (1933).
5a. Boyes-Watson, J., Davidson, E., and Perutz, M. F., *Proc. Roy. Soc. London,* 191A, 83-132 (1947).
6. Bradley, R. S., *Z. Krist.,* 96, 499-503 (1937).
7. Brummage, K. G., *Proc. Roy. Soc. London,* 191A, 243-52 (1947).
8. Brunauer, S., Emmett, P. H., and Teller, E., *J. Am. Chem. Soc.,* 60, 309-19 (1938).
8a. Clark, G. L., and Leppla, P. W., *J. Am. Chem. Soc.,* 58, 2199-201 (1936).
9. Deryagin, B. V., *Nature,* 138, 330-1 (1936).
10. ———, *Bull. acad. sci.,* U.R.S.S., ser chim., 1153-64 (1937).
11. ———, and Kusakov, M., *Acta Physicochim. U.R.S.S.,* 10, 25-44, 153-74 (1939).
12. ———, ———, and Lebedeva, L., *Compt. rend. acad. sci. U.R.S.S.* 23, 671-3 (1939).
13. ———, and Krylov, N. A., *Akad. Nauk, S.S.S.R., Soveshchanie po Vyazkosti Zhidkostei i Kolloid. Rastvorov,* 2, 52-3 (1944).
14. ———, Strakhovsky, G., and Malysheva, D., *Acta Physicochim. U.R.S.S.,* 19, 541-52 (1944).
15. ———, Kusakov, M. M., and Krim, K., *Acta Physicochim, U.R.S.S.,* 20, 35-48 (1945).
16. ———, and Smolinanskii, M., *Compt. rend. acad. sci. U.R.S.S.,* 54, 137-40 (1946).
17. Elton, G. A. H., *Proc. Roy. Soc. London,* 194A, 259-74 (1948).
18. ———, *Proc. Roy. Soc. London,* 194A, 275-87 (1948).
19. Fuller, G. W., Thesis, Stanford University, 1938.
20. Hardy, W. B., *Proc. Roy. Soc. London,* 86A, 610-35 (1912).
21. ———, and Doubleday, I., *Proc. Roy. Soc. London,* 104A, 25-38 (1923).
22. ———, and Bircumshaw, I., *Proc. Roy. Soc. London,* 108A, 1-27 (1925); also Hardy, W. B., in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. I. pp. 288-305, New York, Reinhold Publishing Corp., 1926.
23. Hardy, W., and Nottage, M., *Proc. Roy. Soc. London,* 112A, 62-75 (1926).
24. ———, and ———, *Proc. Roy. Soc. London,* 118A, 209-29 (1928).
25. Hardy, W. B., *Trans. Roy. Soc. London,* 230A, 1-37 (1932).
26. Harkins, W., Mattoon, R. W., and Corrin, M. L., *J. Am. Chem. Soc.,* 68, 220-8 (1946); *J. Colloid Sci.,* 1, 105-26 (1946).
27. Hatschek, E., "The Viscosity of Liquids," p. 94, London, G. Bell & Sons, Ltd., 1928.
28. Henniker, J. C., *Science,* 109, 286 (1949).
29. ———, *Rev. Modern Phys.,* 21, 322-41 (1949).
30. Hess, K., and Gundermann, J., *Ber.,* 70B, 1800-14 (1937).
31. ———, Philippoff, W., and Kiessig, H., *Kolloid-Z.,* 88, 40-51 (1939).
32. Hofmann, U., and Bilke, W., *Kolloid-Z.,* 77, 238-51 (1936).

33. Hughes, E. W., Sawyer, W. M., and Vinograd, J. R., *J. Chem. Phys.*, **13**, 131-2 (1945).
34. Iball, J., *Science*, **109**, 18 (1949).
35. Joris, G. G., and Taylor, H. S., *J. Chem. Phys.*, **16**, 45-51 (1948).
36. Karush, F., and Siegel, B. M., *Science*, **108**, 107 (1948).
37. Kiessig, H., and Philippoff, W., *Naturwissenschaften*, **27**, 593-5 (1939).
38. Kistemaker, J., *Physica*, (s-Grav.) **13**, 81 (1947), Communication #271b from the Kamerlingh Onnes Laboratory of the University of Leiden.
39. Kusakov, M. M., and Krim, K., *Acta Physicochim, U.R.S.S.*, **20**, 548-66 (1945).
40. Langmuir, I., *J. Am. Chem. Soc.*, **39**, 1848 (1917).
41. Lenher, S., *J. Chem. Soc.*, 272-81 (1927).
42. Lennard-Jones, J. E., and Dent, B. M., *Proc. Roy. Soc. London*, **121A**, 247 (1928); *Trans. Faraday Soc.*, **24**, 92 (1928).
43. Levine, S., *Trans. Faraday Soc.*, **44**, 833 (1948).
44. McBain, J. W., and Lee, W. B., *Proc. Roy. Soc. London*, **113A**, 606-20 (1927); see also McBain, J. W., and Alexander, J., in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. III, pp. 9-26, New York, Reinhold Publishing Corp., 1931.
45. McBain, J. W., and Lee, W. B., *J. Phys. Chem.*, **31**, 1674-80 (1927).
46. MacEwan, M. C., *Nature*, **160**, 511 (1947).
47. McHaffie, I. R., and Lenher, S., *J. Chem. Soc.*, **127**, 1559-72 (1925).
47a. Neale, S. M., *Phil. Mag.*, (7) **34**, 577-88 (1943).
48. Palmer, K. J., *Cold Spring Harbor Symposia Quant. Biol.*, **8**, 94 (1940).
49. Schmitt, F. O., and Bear, R. S., *J. Cellular Comp. Physiol.*, **17**, 355-61, 385-94 (1941); **18**, 31-47 (1941); Palmer, K. J., Schmitt, F. O., and Bear, R. S., *J. Cellular Comp. Physiol.*, **17**, 355-67, 385-94 (1941); **18**, 31-47 (1942).
50. Merrill, R. C., and Ballantyne, M., *J. Am. Chem. Soc.*, **70**, 570-7 (1948).
51. Philippoff, W., not yet published.
52. Quincke, G., *Pogg. Ann.*, **137**, 402-14 (1869).
53. Rayleigh, Lord, *Proc. Roy. Soc. London*, **47A**, 364 (1890).
54. Rothen, A., *J. Biol. Chem.*, **163**, 345-6 (1946); **167**, 299-300 (1947).
55. ————, *J. Biol. Chem.*, **168**, 75-97 (1947).
56. Schofield, R. K., *Trans. Faraday Soc.*, **42**, 219-25 (1946).
57. Shaw, B. T., *J. Phys. Chem.*, **46**, 1032-43 (1942).
58. Stewart, G. W., *Chem. Revs.*, **6**, 483-501 (1929); cf., reference 62.
59. Tausz, J., and Szekeley, P., *Erdöl u. Teer*, **9**, 331-3 (1933).
60. Taylor, A. M., and King, A., *Nature*, **132**, 64 (1933); *J. Optical Soc. Am.*, **23**, 308-12 (1933).
61. Verwey, E. J. W., and Overbeek, J. T. G., "Theory of the Stability of Lyophobic Colloids," p. 89, New York, Elsevier Publishing Co., Inc., 1948.
62. Warren, B. E., *Phys. Rev.*, **44**, 969-73 (1933).
63. Zisman, W. A., *J. Colloid Sci.*, **1**, 513-38 (1946); Bigelow, W. C., Pickett, O. L., and Zisman, W. A., *J. Colloid Sci.*, **1**, 513-38 (1946).
64. Zocher, H., and Coper, K., *Z. physik. Chem.*, **132**, 295-302 (1928).

LAWS OF SURFACE TENSION AS EVIDENCE OF AGGREGATES IN LIQUIDS*†

George Antonoff

Introduction

The molecular status of liquids has always been an enigma. The dominant theory of van der Waals assumed that normally a liquid has the same molecular weight as its vapor. Yet certain liquids (e.g., alcohols) behave somewhat abnormally and are believed to be associated. All literature on this subject up to 1915 can be found in a monograph of Turner,[45] which gives a number of empirical rules, but points out that it is futile to accumulate such rules without a comprehensive theory including a general law analogous to Avogadro's law for gases. In Cambridge, after previous experimental work on surface tension, Antonoff developed a theory, with the help of Nicholson, which ultimately led to such a general law for liquids. The study of critical regions proved to be particularly important.

Critical phenomena are found in systems of two kinds: (1) liquid systems which separate into two liquid phases below their critical point of dissolution (the *consolute point*), hereinafter called liquid-liquid systems, e.g., the system phenol-water; and (2) pure substances which separate below their critical point into two phases, liquid and vapor, hereinafter called liquid-vapor systems. In the former case the dissolved substance condenses within a solvent, giving rise to two liquid layers. In the latter case there is no solvent, and condensation takes place, so to say, *in vacuo,*

---

* On January 6, 1950, the author discussed this subject with Dr. Albert Einstein at Princeton, N. J., who wrote him on January 8 as follows:

"Our conversation last Friday has been very interesting for me. You have convinced me that your basic formula $\gamma_1 - \gamma_2 = \gamma_{12}$ deserves attention and careful investigation. In the time of the older van der Waals when the complicated nature of the molecular interactions responsible for the liquid state was not yet known, the general validity of the formula could hardly be doubted. But even now it represents the simplest possibility for a relation between γ_1, γ_2 and γ_{12} for two liquid faces *in a state of mutual thermodynamic equilibrium.* You have also convinced me of the importance of the latter restriction for the possible validity of the formula . . ."

† The subject matter of this paper was the topic of an invited paper read by the author at the Madison meeting of the American Physical Society on June 21, 1948.

ıesulting in the formation of two phases, liquid and vapor. In liquid-liquid systems both phases can be fully investigated experimentally, but this is not possible in liquid-vapor systems because methods of measuring surface tension of vapor are lacking.

The facts mentioned form the basis of the present paper, which stresses the following points: (1) the properties of liquids as a function of temperature are discontinuous, showing kinks at intervals. Densities are especially considered, because they can be measured with extreme accuracy. (2) The properties of liquid-liquid systems are suitable for establishing Antonoff's law, namely, *the value of interfacial tension* (γ_{12}) *between two phases in equilibrium is equal to the difference between the surface tension of one phase* (γ_1) *and that of the other* (γ_2). (3) The theory of Antonoff's law indicates that kinetically such equilibrium is possible only *if both layers contain an equal number of particles per unit volume.* This law is analogous to that of Avogadro, though not identical with it. Within liquids the gas laws are valid: the molecular forces, being balanced, cancel each other. Apparent deviations from the simple gas laws, observed in liquids, are due to·the formation of aggregates. (4) Both phases of a system in equilibrium have the same *colligative properties,* i.e., properties which depend on the *number of particles per unit volume and not on their nature.* Thus two such phases (1) boil at the same temperature, (2) have the same freezing point, (3) the same pressure and composition of vapor, and (4) are isoosmotic, which explains why they do not mix. (5) Liquid-vapor systems are subject to the same laws as the liquid-liquid systems, except that there is no solvent. Both systems display intense opalescence around the critical point. Both are subject to the law of rectilinear diameter.* Antonoff's law is valid in the critical region in both systems. It is thus evident that *liquid and vapor in equilibrium contain an equal number of particles per unit volume.* The association factor, x, is equal to the ratio of the density of the liquid, d_l, and that of the vapor, d_r.† The density changes with temperature discontinuously, as mentioned above, and the aggregation takes place in several stages. Complexity of the aggregates increases at low temperatures, where liquids approach the colloidal state. Changes in the extent of aggregation are responsible for the fluctuations observable when systems are on the way towards equilibrium. Evidence of x-ray analysis supports the theory of aggregates. (6) *All* liquids are subject to the same laws and are *associated.*

* $\dfrac{d_l + d_r}{2}$, plotted against temperature, gives a straight line, called the rectilinear diameter.

† The vapor is assumed to be unassociated.

I. Discontinuities in the Properties of Liquids as a Function of Temperature

Densities. In what follows, the density, d, always means the difference between the density of a liquid, d_l, and that of its vapor, d_v; that is,

$$d = d_l - d_v$$

This is important, since d is a factor in the calculation of surface tension.

Kinks in Density Curves. The exponential function is best used to demonstrate this phenomenon mathematically,* but the calculations are somewhat tedious because three constants must be evaluated.[1] The given constants in equations (1) to (6) below were calculated with figures taken from the work of Young,[3] which are substantially identical with those of Ramsay. The following form of equations apply to all liquids investigated by Young

$$Ae^{kd} = T + A$$
$$A_1 e^{k_1 d} = T + B_1$$
$$A_2 e^{k_2 d} = T + B_2 \ldots$$

where e is the basis of natural logs; A, B, and k (with different subscripts) are constants; and T is the temperature measured in the centigrade scale, counted from the critical point downward.

The theoretical deduction of these equations, as well as an outline of the theory of rectilinear diameter, are given by Antonoff (reference 2, p. 2422), who advanced the view that the thermal expansion of liquids is not linear because aggregated units dissociate on heating. Inserting Young's figures [3] for hexamethylene in the above general equations, we derive:

(1)	$663.0e^{0.8258d}$	$= T + 1000.1$	$(274 - 280°\ T)$
(2)	$24.99e^{3.351d}$	$= T + 77.31$	$(255 - 274°\ T)$
(3)	$18.20e^{3.665d}$	$= T + 52.93$	$(185 - 255°\ T)$
(4)	$7.591e^{4.743d}$	$= T + 18.71$	$(115 - 185°\ T)$
(5)	$2.328e^{6.961d}$	$= T + 5.20$	$(55 - 115°\ T)$
(6)	$0.0727e^{20.97d}$	$= T + 0.0727$	$(0 - 55°\ T)$

Each equation is valid between the approximate ranges of temperatures given at the right. The change (6) starts from the critical temperature which is taken as zero, but which is 280° on the centigrade scale. The constants of these equations are *widely different.* By taking logs of the equations, d becomes a rectilinear function of the expression containing $\log (T + B)$. This is shown in Figure 1, on the basis of equations (2), (3), and (4) above. The angle of intersection of the straight lines is in many cases quite appreciable. This phenomenon cannot be attributed to experimental error. Since the temperature coefficient $\dfrac{\partial d}{\partial T}$ is small, the

* The results given below will be entirely convincing only to those who recalculate them.

accuracy of Young's experiments was ample enough to indicate the above effects.*

Table 1 illustrates equation (3), where the same expression has been logarized twice. The antilog for log d gives the value of d calculated. These values agree perfectly with d experimental, except near the kinks.
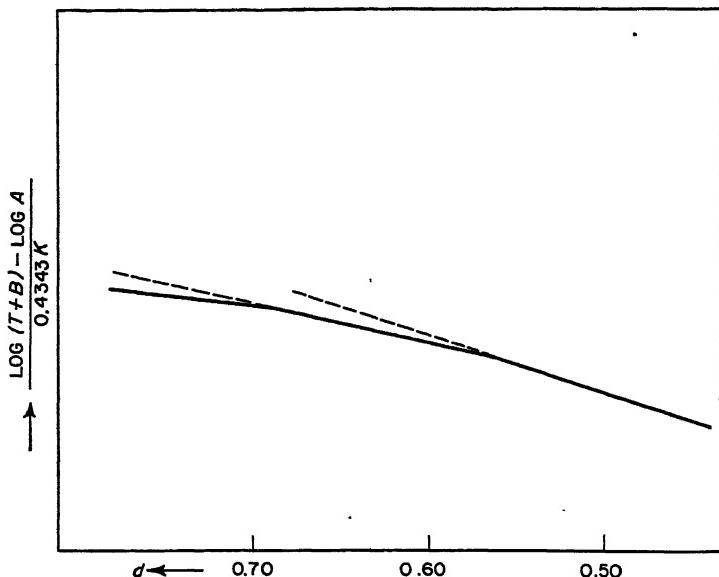

Figure 1. Kinks in density-temperature curve made rectilnear by using loparithmic values.

TABLE 1

$$\log d = \log [\log (T + 52.93) - 1.25990] - 0.20190$$

T	t	d exp.	d calc.
160	120	0.6711	0.6711
150	130	0.65805	0.6581
140	140	0.6442	0.6442
130	150	0.62975	0.6297
120	160	0.6143	0.6143
110	170	0.5981	0.5981
100	180	0.5805	0.5809

This table gives an example of figures between two kinks for hexamethylene (reference 2, p. 2427).

Kink at about 100T. (T is temperature in degrees C, counted from the critical point; t is temperature in degrees C.)

See Antonoff,[1] who also discusses other properties where *density is a factor*. Discontinuities were first observed by Geissler.[4]

Discontinuities in Other Properties. *Latent Heats of Vaporization.* Apart from densities, the same kinks can be observed even to a greater

* This fact was revealed to the author through a lengthy correspondence with Young.

extent in latent heats of vaporization,* which follow the same law (L like d is in the exponent):

$$Ce^{kL} = T + B$$

where L is the latent heat of vaporization; and C, k and B are the constants characteristic for a given interval of the curve between two kinks
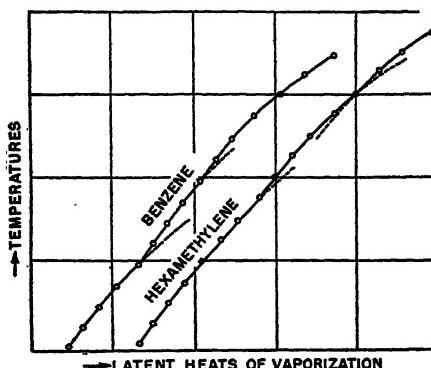

Figure 2. Kinks in curve for latent heats of vaporization.

(Figure 2). These data are reliable because they are derived from density measurements. The drawing shows the kinks distinctly.

Viscosities. For viscosities, the expression will be (T is exponential):

$$\eta + B = Ce^{kT}$$

It is obvious that an accurate exponential expression can be given for $\eta + B$ but not for η.

Surface Tension. The same type of expression can be used for surface tension,[4a]

$$\gamma + B = Ce^{kT}$$

(see Figure 3) which is valid with the given constants only between two kinks.[1]

Concluding Remarks. These formulae work accurately from the critical point downward. However, the procedure of Ramsay is not altogether satisfactory, because it omits the range of temperatures immediately following the critical point, which is the most important part of the curve. If association takes place on condensation, the effects will be most conspicuous in this region. Careful consideration shows that the rest of the curve is not a straight line, as assumed by Ramsay.

Viscosities and surface tensions exhibit the same kinks as densities,[1] although they cannot be measured with the same accuracy as the latter.

* Young calculated these data from the formula of J. E. Mills (reference 3, p. 410), involving densities raised to the power 0.3. This makes the curvature more pronounced, and the kinks become more evident.

These experiments, made from the critical point downwards, have to be accompanied by accurate measurements of densities of the liquid and its saturated vapor. The conclusions given here are based on the density measurements of Young. Antonoff limited his work to the range of temperatures easily attainable in a thermostat. The results showed that all liquids are subject to the same law and *exhibit kinks* at intervals. By way of illustration, a detailed description for propyl alcohol is given [5] in Figure 4, where the kink observed coincides with that indicated by Young's figures. Water was also made an object of special study because it was not included in Young's work, and the experiments are now in progress.[5a]
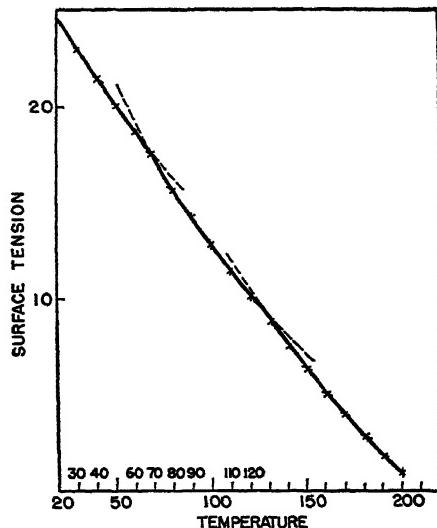
Figure 3. Kinks in surface tension-temperature curve for methyl formate, based on Ramsay's determinations.

For accurate temperature control a mercury-toluene regulator, sensitive to within 0.002°C, was used.[46] Figure 5 shows the regulator as modified by Antonoff. The platinum wire touches the mercury in the capillary intermittently. This contact works well while the mercury is fresh, but when it oxidizes, the contact must be renewed. On raising the temperature, mercury flows out of the capillary, and must be replaced by new mercury from rervervoir *A*. While the excess of mercury covers the capillary, the desired temperature is adjusted. By lowering *A*, the level of mercury in the capillary is fixed for the required temperature. When the platinum wire becomes "stale," it must be heated in a Bunsen flame in order to renew the contact. A Beckmann thermometer, showing fluctuations within 0.002°C, is placed alongside an ordinary thermometer.

The mercury must be renewed periodically, otherwise the accuracy falls to about 0.01°C. Since a temperature difference of 0.1°C would result in an error of only 0.0001 in the density, and since the temperature control never falls below 0.01°C, the kinks observed [5] cannot be due to temperature fluctuations. It should be noted that the kinks are shown in the *third decimal,* whereas the densities can be measured to six decimal places. Antonoff's technique calls for the use of a 25-ml pycnometer and a precision analytical balance.
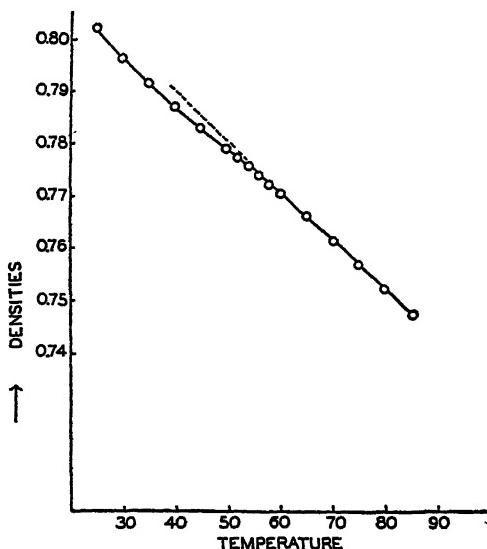
Figure 4. Kink in density-temperature curve for propyl alcohol, observed by Antonoff. One can see that the curve has two branches with different curvatures. The experimental points fit into them with extreme accuracy, except at their intersection (kink), where as a rule a few points are somewhat off the curve.

Distinct evidence of discontinuities can be found in Communications 131, 145, 162, 172, etc. of the Physical Laboratory, University of Leiden, as reported in 1932 by Antonoff (reference 2, p. 2430). The authors observed deviations from the rectilinear diameter, varying between 0.5 and 2 per cent, which is well above experimental error. These deviations are a manifestation of discontinuities in density curves. Even well-demonstrated facts are often rejected when they fall outside the scope of accepted views. Thus, the experimenters at Leiden failed to interpret their own excellent observations. Young likewise failed to notice the facts concealed in his

accurate data.* As pointed out above, the data with hexamethylene show the discontinuities very distinctly. Because the density determinations were made at 10-degree intervals, the kinks, which often occur at intermediate temperatures, are not always evident on a plotted curve. The curve is best drawn on a scale which represents the fourth decimal by a fraction of a millimeter. A flexible ruler backed by lead can locate the kinks which coincide with those found mathematically. On such a scale, the kinks are in most cases visible to the naked eye.†
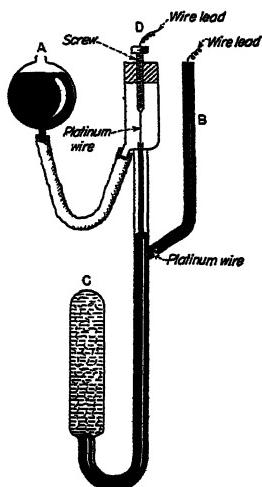
Figure 5. Mercury-toluene thermoregulator. Reservoir A is filled with pure mercury; tube B is filled with mercury to establish contact; C is filled with toluene and mercury; D is connected with a screw to adjust the contact with mercury in the capillary.

II. Surface and Interfacial Tension

Interface Liquid-Liquid. *Antonoff's Law.* The relationship between two liquid layers *in equilibrium with each other,* known as Antonoff's law,[6] was formulated in 1907:

$$\gamma_{12} = \gamma_1 - \gamma_2$$

where γ_{12} is the interfacial tension, γ_1 the surface tension of one layer, and γ_2 that of the other,‡ both measured against their common vapor. Special methods evolved to demonstrate this law are described below.

Causes of Difficulties. Two problems confronted those who tried to test the validity of Antonoff's law. First, it is very difficult to attain

* On July 9, 1924, Young wrote the author as follows: "There is, of course, no doubt that your calculations appear to indicate discontinuity in the case of benzene. . . . There can be little doubt that with increased experience and improved methods, the data for paraffins, hexamethylene, etc., are more accurate than the earlier data."

† A. Urmanczy of Oslo (Norway) has just advised that his work, soon to be published, establishes the existence of kinks in several liquids.

‡ Strictly speaking, γ_1 and γ_2 are interfacial tensions between a liquid and vapor. At low temperatures, when the surface tension of vapors is negligible, the figures obtained can be considered surface tensions of liquids. These terms are often used indiscriminately.

equilibrium,[7] which is not always reached instantaneously in liquids, but may take days, weeks, and even months.[8] The other difficulty involves experimental methods. The capillary method is generally believed to be most reliable,[9] although in many cases it requires special handling.

In systems consisting of two liquid layers, one layer is generally polar and the other nonpolar. The predominantly polar layer causes difficulties, especially when the mutual solubility is small and one or both components are volatile. Under these conditions it is not very easy to maintain equilibrium within the narrow capillary and to arrive at reproducible results. To diminish frictional effects within the capillary, short capillaries were used.[9a] *The capillary rise was found to be dependent on*
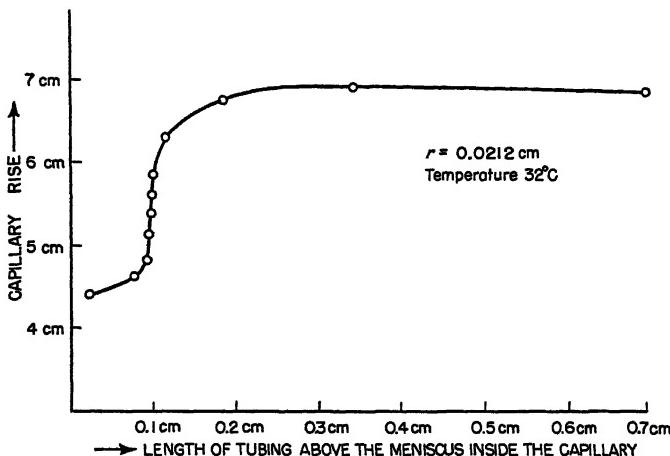
Figure 6. Dependence of capillary rise on the length of capillary.

the length of the capillary above the meniscus. The shorter the capillary, the flatter the meniscus becomes. Thus, in capillary measurements the length of the capillary must be specified. In order to obtain the maximum value of the capillary rise, in the case mentioned above, it is necessary to extend the capillary several centimeters, i.e., 3 to 4 centimeters, above the meniscus within it. Further lengthening of the capillary has no additional effect. Since the value of $\frac{d\gamma}{dc}$ in the aqueous layer is enormous (γ = surface tension, c = concentration), a part of the above effect may be attributed to change of concentration due to evaporation at the surface.

Experiments with pure water, made by Conan in the author's laboratory at Fordham University, showed a dependence of the capillary rise on the length of the capillary above the meniscus in the tube, the

optimum length being about 0.2 centimeter (see Figure 6). This phe-
nomenon does not follow classical theory. The length, 0.2 centimeter,
is well above the radius of molecular action beyond which molecular
forces are not supposed to be effective.

According to present evidence, polar layers exhibit this effect, while
nonpolar layers, such as water-saturated benzene, do not show it ap-
preciably. Capillary rise at the water-ether interface definitely showed
dependence on the length of the capillary.*

Role of the Solid and Angle of Contact. As previously shown [10] with
solutions of isobutyric acid of various concentrations, the solid, i.e., the
material of the capillary, must have a much higher surface tension than
the liquid. If the solution has a higher surface tension than the solid,
there is capillary depression, as in mercury-glass; in the reverse case there
is capillary rise; and when the surface tension of the solid is equal to
that of the liquid, there is neither rise nor depression.

Antonoff's law is valid at the solid-liquid interface, as will be discussed
later. In this situation it can be expressed as

$$\gamma_s - \gamma_l = \gamma_{sl}$$

where γ_s is the surface tension of the solid, γ_l that of the liquid, and γ_{sl}
is the interfacial tension. If $\gamma_s > \gamma_l$ there is wetting and γ_{sl} is positive.
If $\gamma_s < \gamma_l$ there is no wetting and γ_{sl} is negative, as in the case of mercury-
glass. Solids differ in surface tension. Those with low surface tensions are
lyophobic. The materials having a high surface tension are *lyophilic.*

When $\gamma_s = \gamma_l$ there is no capillary rise or depression and the meniscus
in the capillary appears flat. When γ_s increases in value the meniscus
begins to curve, but complete wetting is attained only when γ_{sl} acquires
a large value. Only when the angle of contact becomes zero is the cap-
illary method suitable for measuring surface tension.[11]† In order for the
angle of contact to become zero, a pull from above is necessary, and it
comes from the material of the capillary above the meniscus within it.

Capillary rise is explained as follows: The force of attraction between
the solid and the liquid is greater than that between the particles of
the liquid itself. The film of liquid fixed on the wall of the capillary is
capable of supporting a column of liquid, whose height is determined
by the surface tension of the liquid. The weight of this column is *not*

* To give a full explanation of these effects and the peculiar phenomena that have
been observed, a research of long duration would be required, which is beyond the
scope of the present paper. These facts are mentioned to explain why the capillary
method was used with caution.

† Antonoff has never observed a contact angle that equalled zero, except the mo-
ment after the capillary was lifted a little. Viewed from the telescope of the cathetome-
ter, the meniscus descends for some time; then the angle of contact ceases to be zero.
Moreover, the capillary behaves quite differently when it is previously dried or kept
in a liquid before use. He therefore does not recommend precision measurements by
the capillary method.

equal to $2\pi r \gamma_l$ (r being the radius of the capillary). The thickness of the immobilized film may vary with the range of force exercised by the solid, and must be deduced from r. The effective radius can be determined either by emptying the capillary and weighing the liquid, or by weighing the capillary when dry and when wet.*

Owing to the difficulties described above, the capillary method was not favored in these experiments. In systems with very small mutual solubility, the drop method was used for the layer with the higher surface tension (in most cases, the aqueous layer). A pipet of any kind can be
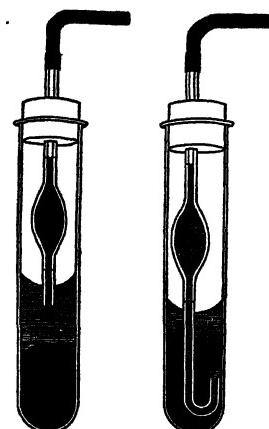
Figure 7. Method of discharging drops of one liquid within another, for measuring the interfacial tension. *Left:* Heavier chloroform layer discharged in the aqueous upper phase; pipet not curved. *Right:* Lighter benzene layer discharged in the heavier aqueous phase; pipet curved.

used; the pure water discharged from it first and the saturated solution afterwards. Quincke† recommended the formula

$$\gamma = \gamma_0 \frac{dm_0}{d_0 m}$$

where γ_0 is the surface tension of water, d_0 its density, and m_0 the number of drops. The same symbols, without subscripts, refer to the saturated solution. This formula is valid for systems with very small mutual solubility.

The measurement of interfacial tension is also difficult. According to Quincke, it is necessary to moisten the capillary with the liquid of higher surface tension first. However, the meniscus is liable to acquire an arbi-

---

* For this reason it appears that the classical concept of the angle of contact has outlived itself. It is the general curvature some distance from the wall that matters. This is a function of surface tension. It can be better observed in broader tubes. By suitable calibration, the surface tension could be expressed in terms of the curvature. Below it is shown that a flat surface means that the surface tension is zero. The difficulty lies in finding a suitable standard. The wave method of measuring the surface tension might be applicable because the measurement is not influenced by the proximity of any solid.

† The author does not give reference, as he worked in Quincke's laboratory and made citations from his own notes.

trary position within the capillary, unless the latter is very broad. The radius of the capillary was, as a rule, about 0.02 centimeter, unless otherwise specified.[12] The friction is probably due to emulsification, which in turn is apparently caused by slight local thermal fluctuations. The emulsification becomes visible when two layers remain standing for a certain length of time. Good results can be obtained only with fresh surfaces, either by the drop method or by the pendant drop method.
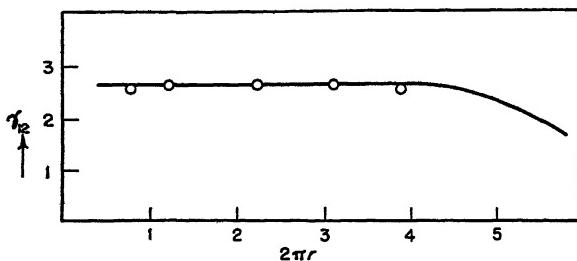

Figure 8. Dependence of interfacial tension measurements on internal diameter of the pipet outlet.

Using the drop method for measurement of interfacial tension, the liquid with lower surface tension must always be discharged into the one with higher surface tension.* In this way, the drop will not wet the external part of the pipet (Figure 7). The results are independent of the diameter of the tip, within wide limits. It was found that γ_{12} can be derived from the relationship $\gamma_{12} = \dfrac{P}{2\pi r}$, where P is the weight of the drop,
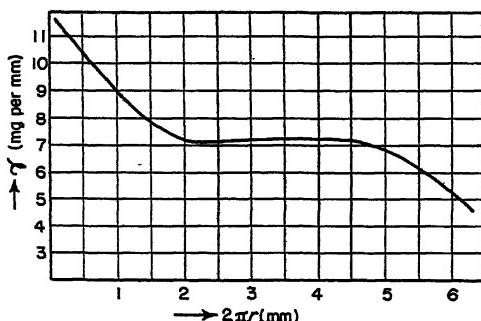

Figure 9. Dependence of surface-tension measurements on the thickness of the wall of the pipet outlet.

and $2r$ the internal diameter of the tip (Figure 8). This conclusion was arrived at by discharging drops of water from pipets of different diameters, with extremely thin-walled orifices (Figure 9). There is a wide region where $\gamma = \dfrac{P}{2\pi r}$, and is equal to the values of surface tension

* The speed of discharge must be controlled.

given in tables of physical constants. The process of drop formation is the same as in the previous example; the internal diameter of the orifice is the dominating factor.

Since thin-walled pipets are brittle and difficult to make, the above method was used only for calibration. The pipets were prepared as follows: A tube of soft glass, closed at one end, was inserted in a wider tube of Jena glass, which was heated by a blow pipe, so that the inner tube was softened by uniform heat. It was then drawn out while air was being blown into it. The tube prepared in this manner may have walls so thin that they exhibit rainbow colors, but it is so uniform that it can be evacuated wtihout collapse.* The tube is embedded in paraffin (which is later removed by xylene), and then cut across with a razor. The remaining end of the tube was used to measure the diameter of the tube microscopically or otherwise.

Experimental figures obtained by Conan in the author's laboratory, using these methods, are given below for some systems which presented great difficulty owing to small solubility, volatility, or both. All chemicals used were c.p. products.

Benzene-Water at 25°C. The system was kept in a thermostat for about ten days, during which time the periodic changes that preceded equilibrium were observed. The figures were calculated by a good slide rule, the accuracy of which was equivalent to three-figure logs.

Aqueous Layer γ_1	Benzene Layer γ_2	γ_{12}	$\gamma_1 - \gamma_2$	diff.
56.52	29.25	27.20	27.22	0.02 dynes/cm

If the pendant drop method had been used with equilibrium conditions, the results would probably have been the same, because both methods are expressed in terms of the capillary height method for water.[18]

Concordant results can be obtained for the aqueous layer also by the capillary method, if the system is first brought into a state of equilibrium by prolonged standing in a thermostat. Thus, the same surface tension figure for the aqueous layer was obtained by this method as by the drop method above. No corrections are needed here, as the standard values for pure water were also without correction.

The difficulty of measuring the separate tensions of the aqueous phase was pointed out by Carter and Jones,[14] who used the maximum bubble-pressure method. However, their figures have no definite meaning since they ignored the conditions of equilibrium. The work of Harkins [15] is open to the same criticism. In no way does it reflect on the validity of Antonoff's law.

---

* The author learned this method from Baumbach, the glass blower of Manchester University, who used it for making emanation tubes for Rutherford, which were permeable to alpha particles but not to helium.

Chloroform-Water at 25°C. In this system no equilibrium can ever be attained as chloroform undergoes hydrolysis.

Days after Mixing	Aqueous Layer γ_1	Chloroform Layer γ_2	γ_{12}	$\gamma_1 - \gamma_2$	diff.
3	51.20	27.14	21.77	23.79	2.02
10	51.65	27.41	22.99	24.24	1.25
16	49.45	27.41	22.21	22.04	0.17

A similar behavior has been observed in the system CS_2 + water, where an irreversible change makes equilibrium impossible.

Ether-Water at 25°C. The following figures were obtained for this system.

Days	Aqueous Layer γ_1	Ether Layer γ_2	γ_{12}	$\gamma_1 - \gamma_2$	diff.
0	29.42	17.98	7.70	11.44	3.74
3	28.38	17.53	8.06	10.85	2.79
7	26.82	17.39	8.14	9.43	1.29

Measurements were resumed a month later. Marked changes were then observed, probably owing to the presence of oxidation products.

Conclusion. Provided there are no irreversible changes, the methods described above can give reproducible results. Special purity is of no importance, because impurities help to establish equilibrium. This can be demonstrated by adding substances in varying quantities. However, impurities are of great importance to those studying the specific properties of surfaces, which is an entirely different problem.

All experimental data given in this section rest on the assumption that the surface tension of water at room temperature is about 72 dynes per cm. The figures thus obtained consistently indicate the definite relationship $\gamma_{12} = \gamma_1 - \gamma_2$, even though they may not be the *true values* of surface tension.* Systems exhibiting appreciable mutual solubility present no difficulties, and results conformable with Antonoff's law have been obtained for a number of such systems, provided they are brought into a state of equilibrium first. Quincke and Rayleigh failed to develop surface tension relationships because equilibrium was not considered. Experiments in which one pure liquid is discharged into another are valueless because equilibrium does not exist at the time of measurement. These results do not seem to corroborate Antonoff's law absolutely (except in the case of the benzene-water system), but they cannot be interpreted as evidence against it, since Antonoff's law is applicable only to systems in equilibrium.

The chief experimental difficulty is the peculiar behavior of matter in narrow capillaries. The belief was even expressed that the surface tension

* The use of other methods for determining the surface tension, γ, gives the same relationship.

of a liquid in a narrow capillary is abnormally great.[16] All these problems arise from the enormous fluctuations taking place in the liquid layer with the higher surface tension; but they disappear when, on standing under fixed conditions, the system reaches equilibrium. These fluctuations also exist in other systems, and have been observed in the aqueous solutions of isobutyric acid, the solutions of amylene in aniline, etc. The disturbances appear to be in the nature of pulsations restricted to a small volume, as they are observed within narrow capillaries. The same substances, when discharged dropwise in an atmosphere of saturated vapor, do not show fluctuations, for the effect appears to be averaged out. Thus, very substantial changes take place within a system when it reaches a state of equilibrium.

Interface Liquid-Solid. At the liquid-solid interface, the same form of relationship holds true as for the liquid-liquid interface. If there is no interaction between the solid and the liquid, the relationship

$$\gamma_{sl} = \gamma_s - \gamma_l$$

is established at the time of contact. This concept provides a method for measuring the surface tension of solids. Of the three magnitudes, only one can be measured by known methods. If the value γ_l can be measured, it can be made to equal γ_s, when γ_{sl} becomes zero. This status can be attained with certain pastes whose surface tensions may vary between 30 and 900 dynes per cm. If the paste is too stiff, it can be warmed. The temperature coefficient of surface tension is about 0.5 per cent, whereas that of viscosity may be as large as 25 per cent. If the paste wets the solid, $\gamma_s > \gamma_l$; if it does not, $\gamma_s < \gamma_l$. At the transition point $\gamma_s = \gamma_l$. The accuracy of this measurement depends on the number of pastes of different concentrations prepared for the experiment. Details are given in journal references.[10, 12]

The theoretical aspect of surface tension of solids has been emphasized by Gibbs and Curie. The subject has recently been reviewed by von Laue,[17a] who gives a geometrical interpretation of the theory under the name of Wulff's law: "If the free energy of the surface of a crystal has a minimum value, there must exist within it a point (*Wulff's point*), situated at such distances from the surfaces, that the ratio of the distances is equal to the ratio of the surface energies." The methods of measuring the surface tension of solids were not known to the above authors. Wulff's law can now be experimentally substantiated. This work is being done, using Antonoff's methods, by Van Hook, at Holy Cross, who finds widely different values of surface tension for different crystal faces. It appears possible to give a physical interpretation of the above law: The structural units are differently situated along each plane. They meet at distances whose ratio is equal to that of their surface tensions.

III. Theory of Surface Tension and Its Consequences

Surface Tension and Its Correlation with Internal Pressure. When Laplace originated the theory of surface tension, nothing was known about the structure of atoms and molecules. To explain the phenomenon of surface tension he had to make a hypothesis *ad hoc.* The "classical concept" assumes that the molecules are spheres exercising a uniform field of force over a very short range. The whole picture can be simply presented, without involving the law of molecular attraction, by using a crystal as a model (Figure 10). Let us assume that the force of attraction
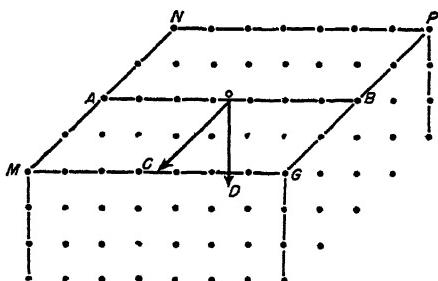
Figure 10. Crystal model, each dot representing a molecule. The arrows indicate correlation between surface tension and internal pressure.

is effective between two adjacent molecules, but that it vanishes so rapidly with distance that its effect on more distant molecules is negligible. Let AB be a line of unit length on the surface. The row of molecules along the line AB will be under the influence of a force tangential to the surface, *which is called the surface tension* (γ). A similar force acts on AB downwards. If the adjacent rows have no effect, the inward force per row will be equal to the tangential force. The total inward force will be equal to the sum of the effects of all rows. This inward force is called the internal or molecular pressure (P). Thus, if n is the number of molecules per unit volume, their number per unit length will be $n^{1/3}$, and per unit area, $n^{2/3}$. And the relationship may be expressed as

$$P = K\gamma n^{1/3}$$

The proportionality factor, K, equals one for the model with the uniform field of force illustrated in Figure 10.

The same theory and type of equation can also be applied to liquids. However, as liquid particles are in constant motion, all magnitudes in the above expression will acquire *statistical values.* This relationship is equally valid for all three states of aggregation, although for gases one can only show qualitatively the existence of surface tension (see section V below). The surface tension of liquids can be measured, but since P can-

not be determined experimentally, the above equation cannot be utilized. If it were possible to measure P, one could determine n (the number of particles per unit volume), and thus calculate the degree of association. As regards the solid state, it appears at first sight that P can be calculated from breaking-stress experiments with crystals, but these experiments give very low values [18] because of the invisible cracks which are always present in real crystals. If a crystal is "cured" in a hot saturated solution of the same substance,[19] the breaking stress is greatly increased.* It will be some time, however, before such problems concerning the solid state can be resolved quantitatively, and it is therefore advisable to concentrate on liquid systems. If n cannot be estimated by the formula given in section III, it can be derived indirectly by using liquid-liquid systems in a state of equilibrium. However, only systems consisting of two partially soluble liquids, where the media on either side of the interface are fluids, can be successfully investigated. This is an essential condition for experimentation by known methods.

Antonoff's Law Deduced. From the above considerations, Antonoff's law (see section II) can be deduced theoretically. When two liquids are in equilibrium, the following relationship must be valid:

$$P_{12} = P_1 - P_2$$

where P_1 is the inward pressure in one layer, P_2 that in the other, and P_{12} is the resultant force at the interface. P_1 is expressed in terms of surface tension by

$$P_1 = K\gamma_1 n_1^{1/3}$$

The inward pressure for the other layer, similarly expressed would be

$$P_2 = K\gamma_2 n_2^{1/3}$$

Therefore

$$P_{12} = P_1 - P_2 = K(\gamma_1 n_1^{1/3} - \gamma_2 n_2^{1/3})$$

and

$$P_{12} = K\gamma_{12} n_{12}$$

Two Equimolecular Liquid Layers. The expression derived above becomes identical with Antonoff's law† if

$$n_{12} = n_1 = n_2$$

which means that both layers contain an equal number of *particles* per unit volume. Here γ_1 and γ_2 are the surface tensions of saturated layers

---

* Carefully annealed glass tubes can withstand pressures up to 1000 atmospheres, in a Cailletet pump, if after annealing they are prevented from touching other glass (in the latter case invisible cracks are formed). One can boil water for years in an annealed glass, whereas ordinary glass usually breaks very soon.

† K must be the same in these expressions if the molecular concentrations are identical.

measured against their common vapor; γ_{12} is the interfacial tension; n_1 is the number of particles per unit volume in one layer; n_2 the number of particles per unit volume in the other layer; and n_{12} is the number of particles per unit volume in the immediate neighborhood of the interface. But this can only be true if combinations are formed. The molecules added must combine with some of those in solution without increasing the total number of molecules. These combined or complex molecules will be called *particles* or aggregates, to distinguish them from simple molecules. However, they are not chemical compounds in the ordinary sense because they cannot be isolated. They are "equilibrium compounds," existing only under equilibrium conditions. Experimental evidence fully supports this view.

Cancellation of Molecular Forces Within the Liquids. Internal pressure is an elusive property. It is known that a greased steel needle can float on the surface of water, since it is supported by surface tension forces.
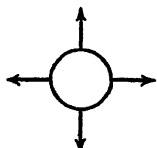
Figure 11. Molecule pulled by equal forces in opposite directions. The forces are balanced and thus cancel each other.

If forced into the water, it sinks, because it is then affected by balanced forces; on the surface, the forces are unbalanced. However, this does not signify that internal pressure does not exist. If a liquid is enclosed in a cylinder with a piston, the work necessary to break the properly deaerated column of liquid is enormous.

Theory of Solutions: Gas Laws. Let us assume that in a substance, *A*, a certain amount of substance, *B*, is dissolved. According to van't Hoff's theory, the dissolved substance behaves like a gas and is subject to all the gas laws. The solvent does not contribute any of its properties to the solute; in fact its influences are negligible.

The gas laws are supposed to be accurate only in dilute solutions. In concentrated solutions deviations are generally observed which are attributed by many authors to van der Waals' attractions. Yet, from the above considerations, it is obvious that these irregularities are not caused by molecular forces,* but rather by the number of molecules in solution. Substance *A* may be associated, containing particles of various degrees of complexity in addition to simple molecules. The same applies to substance *B*. Furthermore, *A* and *B* can also form associated molecules of mixed composition. We are here confronted with a complex system.

---

* In its effect this theory is identical with that of Duclaux, but the interpretation is different.[20]

Modern Views. We now know that atoms and molecules are not simple
spheres, as the classical theory assumed. There is no fixed law of molec-
ular attraction; this factor may vary between the inverse seventh power
and inverse square of the intermolecular distance. In the latter case we
are dealing with long-range forces. However, the results presented here,
relating surface tension and inward pressure, are not dependent on the
law of molecular attraction. The possibility of a long-range force, how-
ever, is an important factor. As a result of optical observation, McBain
believes in long-range forces.[22] It is significant that polar substances in
capillary experiments behave in such a way that it appears necessary to
invoke the existence of long-range forces. Nowadays it is evident that
"no single and simple type of force field represents the results with any
degree of exactness."[23, *]

IV. Experiments Corroborating the Theory

Definition of Equilibrium. Systems are in equilibrium if their prop-
erties do not change with time. In liquid-liquid systems this condition is
not necessarily established by simply *mixing the two liquids.* Apart from
thermal equilibrium established in 20 to 30 minutes, reversible changes
taking place in solutions require time. In some cases the changes are
rapid, in others they take several weeks or months. The latter is true of
substances of a high degree of purity. Impurities act catalytically to
accelerate the attainment of equilibrium. Thus, the time factor must
be introduced into these studies as a definite parameter. The physical
properties of liquids may be entirely different, depending on their pre-
vious history.† For example, for the phenol-water system one author
found [24]

$$\gamma_1 - \gamma_2 > \gamma_{12}$$

Another found:

$$\gamma_1 - \gamma_2 < \gamma_{12}$$

However, neither of them worked under conditions of equilibrium. In
the one case, the phenol used was kept in a cool place; in the other, in
a warm place. It takes from six to ten days to bring the system phenol-
water into equilibrium, after which time the relationship

$$\gamma_1 - \gamma_2 = \gamma_{12}$$

holds consistently. When systems proceed toward equilibrium, their
physical properties fluctuate within wide limits. In most cases the third

* Details regarding the structure of molecules and the nature of forces between them
can be found in P. Debye's monograph, "The Dipole Moment and Chemical Struc-
ture," London, Blackie and Co., Ltd., 1931.
† This is one reason why alcoholic beverages change their taste in time.

decimal is affected in densities, indicating that this is not a *surface* phenomenon. These fluctuations practically vanish when equilibrium is established. In a state of equilibrium there can be no difference of potentials between components.*

Crucial Experiments and Their Consequences. The theory indicated that two liquid layers in equilibrium must contain an equal number of particles per unit volume. How is this possible when the concentrations of the layers are so different? Investigation of the freezing-point curves explains this phenomenon.[25] A typical example is given in Figure 12.
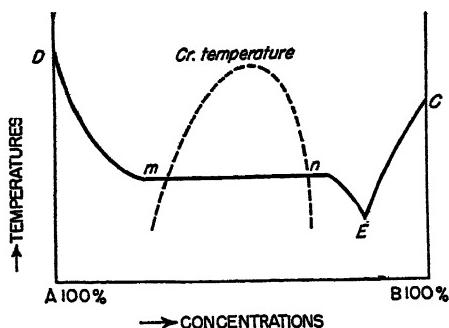
Figure 12. Typical freezing-point curve of two-phase liquid-liquid systems.

For the pure substance, *A,* the freezing point is shown by the highest point on the diagram. The addition of solute, *B,* causes lowering of the freezing point until a separation into two phases takes place. The dotted curve is the so-called solubility curve, within which there are two liquid layers; outsde this curve the system is homogeneous. The curve dips down to point *E,* called the eutectic point. On further increase of concentration, the curve rises again and reaches *C,* the freezing temperature of the component, *B.* At point *m,* a trace of the second layer appears. Its relative volume increases with further addition of component *B.* At point *n,* the first layer disappears and the whole mass is represented by the second layer. Between *m* and *n,* the system consists of two layers and the freezing point remains constant. Thus, the two layers are made up of two solutions situated on one side of the eutectic point; they are therefore solutions in the same solvent and on freezing they separate the same kind of ice.

Colligative Properties Identical for Two Liquid Layers in Equilibrium. The freezing point is one of the four colligative properties (see page 84) which are identical for any two liquid layers in equilibrium. Thus, if the gas laws hold, the two layers contain an equal number of particles per unit volume.

* If two identical electrodes are inserted into a two-layer system in equilibrium, no difference of potential can be observed if the system conducts the electrical current.

Conclusion. No better confirmation of the theory can be expected. The very formation of two liquid layers necessitates the formation of complex molecules in solution. The actual experimental data are given in journal references.[25]

V. Extension of the Theory to Liquid-Vapor Systems

The analogy between liquid-liquid and liquid-vapor systems is striking, except that in the investigation of the latter we are handicapped by the lack of suitable methods. We cannot measure the surface tension of a vapor experimentally, and we have no means of showing directly whether Antonoff's law is valid for liquid-vapor systems. However, a study of critical phenomena will shed some light on the question.

Critical Region. Known critical phenomena are of two kinds:

(1) In some cases a substance condenses within a solvent. This condensation results in the formation of two liquid layers, or two liquid phases, which we describe as liquid-liquid systems. The latter systems are suitable for establishing Antonoff's law. As a rule these systems have a critical point,* called the *critical point of dissolution* or *consolute point,* above which the two layers mix, whereupon the systems become homogeneous. In the critical region, i.e., all around the critical point, the surface tension is independent of concentration;[6] i.e., when separation into two phases takes place, both phases have the same surface tension, as can be shown by direct measurements; γ_{12} equals zero, and the meniscus separating them is *flat.*

(2) In other cases, a pure substance condenses below its critical point in a space not otherwise filled, or, *in vacuo.* It separates into two phases, liquid and vapor. Just below the critical point the meniscus is *flat* and there is no capillary rise. This means that Antonoff's law in its special form

$$\gamma_1 - \gamma_2 = \gamma_{12} = 0$$

is valid, where γ_1 is the surface tension of the liquid, and γ_2 that of the vapor. By the capillary method, we measure the interfacial tension γ_{12} equals zero. This shows that vapors have a surface tension, because the flat meniscus is an indication that both phases have the same surface tension. This procedure is referred to in the literature as Antonoff's method of flat meniscus.[26]

Colligative Properties. In liquid-liquid systems we may make use of the four colligative properties; we base our conclusions as to the molecular status of the dissolved substances on the effect they produce on the

* This region is characterized by an intense opalescence, which can best be observed in the aniline-amylene (trimethyl-ethylene) system, where the critical temperature is a little lower than room temperature.

solvent. The two liquid layers have the same vapor pressure and they are isoosmotic. This can be understood if one assumes that simpler particles in one layer are in equilibrium with more complex particles in the other, both dissolved in the same solvent.

If we imagine that the solvent has vanished, what remains is an exact picture of a liquid-vapor system. The vapor exercises a certain pressure outwardly; so does the liquid. These two pressures must be equal. Pressure is a colligative property because it depends on the number of particles per unit volume.

Rectilinear Diameter. The analogy with liquid-liquid systems is also seen in the so-called rectilinear diameter (Figure 13, line R). In liquid-liquid systems the solubility curve is determined by Alexeiev's method.
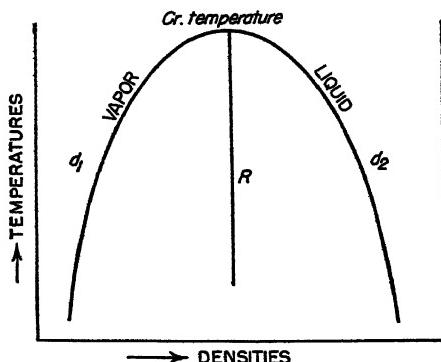
Figure 13. Density - temperature curve comprising two phases. Curve has shape of an asymmetric parabola. Outside it the system is homogeneous; inside it there are two phases. At the apex of the curve is the critical point. Line R is the rectilinear diameter. For liquid-liquid systems, this curve is called the solubility curve.

Liquids of different concentrations are sealed in test tubes. The temperatures are observed when the liquids separate into two layers or otherwise. The temperature-composition curve is a slightly asymmetrical parabola, the two branches of which meet at the apex where the critical point is located. Or one may plot densities against temperatures. The branch on the right gives the densities of the heavier layer (d_2), and that on the left, the densities of the lighter layer (d_1). There is a similar curve for liquid-vapor systems, with d_2 representing the density of the liquid, and d_1 that of the vapor. The values $\dfrac{d_1+d_2}{2}$, plotted against the temperature lie on a straight line, which is known as the rectilinear diameter of Cailletet and Mathias.

Aggregates in Pure Liquids. Assuming Antonoff's law valid for liquid-vapor systems, both *liquid and vapor* must contain an *equal number of particles* per unit volume. The process of association of simple molecules begins well above the critical point. At some distance from the critical point, the pressure-volume curve becomes horizontal, the pressure (P)

becoming independent of volume (V) in the critical region extending all around the critical point, the region being characterized by the phenomenon of opalescence (Figure 14).*

The interpretation of the process of condensation in the critical region [28] is as follows: Simple molecules combine with each other to form double molecules in such a way that by the time the critical point is reached, one-third of the molecules has combined with another third.
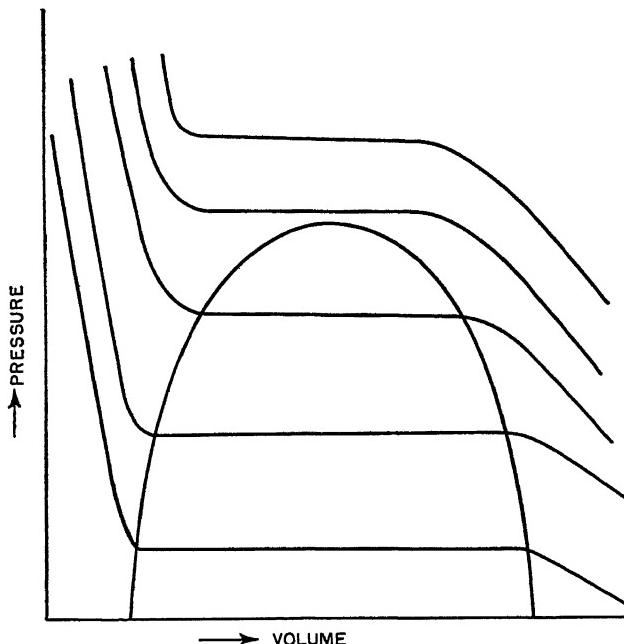

Figure 14. Pressure-volume relationship near the critical point, as obtained experimentally. It differs from diagrams based on van der Waals' equation.

Thus, the system contains an equal number of double and single molecules. Just below the critical point, the two kinds of particles separate without an appreciable expenditure of energy. The separation takes place at a certain temperature because the double molecules suffer a diminution in kinetic energy, which causes their sedimentation. Under these conditions, just below the critical point the ratio of the densities is

$$\frac{d_l}{d_g} = 2$$

Thus, if the molecular weight of the gas is A, that of the liquid in this condition will be A_2. In other words, the association factor is 2. However,

* A theory of opalescence has been given by Einstein and Smoluchowski.[27]

it increases with the decrease in temperature because d_l increases and d_v decreases. At low temperatures, the degree of complexity of the aggregates formed becomes enormous: liquids approach the realm of colloids.

The fact that densities change with temperature in a discontinuous manner suggests that the process of aggregation of matter takes place in several stages. If one estimates the ratio $\dfrac{d_l}{d_v}$ corresponding to kinks, the ratio appears to be equal to some simple integer. However, the existing data are not accurate enough and the conclusions derived are of speculative nature. The following changes were deduced for hexamethylene[2]:

$$
\begin{aligned}
A + A &= A_2 \\
2A_2 &= A_4 \\
\text{or } 3A_2 &= A_6 \\
6A_4 &= A_{24} \\
\text{or } 4A_6 &= A_{24} \\
4A_{24} &= A_{96} \\
6A_{96} &= A_{576}
\end{aligned}
$$

X-Ray Analysis. A similar picture is revealed by x-ray analysis. These results can be found summarized in a statement by Mark.[29]

"Individual particles of a liquid do not, like those of a perfect gas, assume at random all possible positions and orientations in space but their mutual arrangement resembles very largely that of the crystalline state. In this sense we may conceive of the liquid as an aggregate of numerous very small crystals in which each individual does not show exactly the arrangement characteristic for it in the crystal but assumes a series of temporary positions. These small crystallites (each consists only of 10 to 100 atoms) are only very short-lived. They disband continually and recrystallize so that a given particle belongs for an instant to one of these small aggregates, then to none, then to a subsequent one which has been formed in the meantime, and so on. . ."

However, these crystallites are too small to be taken into account in this theory. It is therefore assumed that there must be a secondary structure, similar to that proposed by Bragg [30] for the crystalline state. This structure is not revealed by x-ray analysis. The cybotactic groups of Stewart apparently are a manifestation of this same phenomenon.[31]

All Liquids Are Associated. This theory indicates that at low temperature the liquid must consist of enormous aggregates which are in equilibrium with more simple particles in the vapor phase. The number of particles in both phases must be the same. The mobile equilibrium between liquid and vapor must be maintained in the following way: Large particles are shot off by the liquid into the vapor, where they disintegrate into simple molecules, thus increasing the vapor pressure. This increased pressure causes the return of molecules from the vapor to the

liquid, where they join the aggregates in the manner described. Similarly, we can imagine an equilibrium comprising the solid phase.[32] According to the theory outlined, all substances are associated, irrespective of whether they are described by current theories as normal or associated. Also, the interval of temperature between two kinks may be different in various cases.

Causes of Irreproducibility of Experimental Data. Liquids consist of several molecular species which are not necessarily in equilibrium with
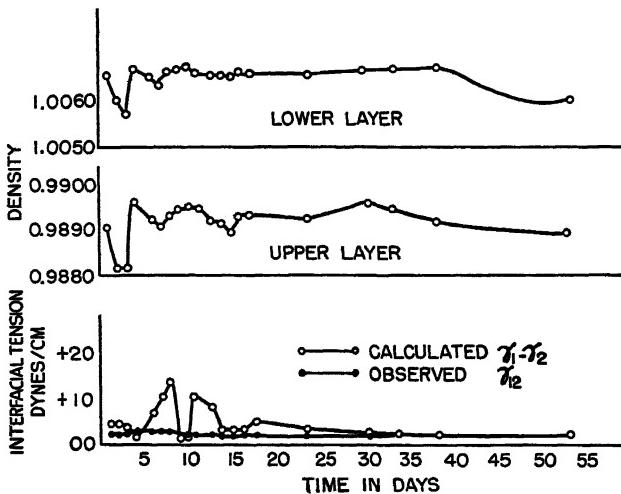
Figure 15. Fluctuations in properties of liquid-liquid systems on the way towards equilibrium.

one another. Impurities, such as water, may have a catalytic effect in the process of attaining equilibrium. The experiments of Baker on excessive drying[33] indicate this. Although his results have not received general recognition, there is certainly an element of truth in them. Experiments with phenol are also of interest in this connection. When freshly distilled, phenol is not hygroscopic, but after six months or so it becomes sticky. Even in a sealed zinc container, phenol generally liquefies after standing for several months or years. This is not due to moisture, but to some reversible process whereby phenol is transformed into another modification, possibly its keto form. The initial content of moisture determines the velocity with which the above transformation takes place. That is why numerical data for liquids are often irreproducible and figures by different authors diverge within wide limits.

Fluctuations. The properties of two-phase liquid systems approaching equilibrium fluctuate. The two layers were kept together in a thermostat

and small portions were taken for analysis. A 25-cc pycnometer was filled, brought to the mark in a thermostat, and weighed. The aqueous layer showed marked changes in density, and the surface tensions of both layers also fluctuated. These changes ceased when the system reached equilibrium, and it was therefore concluded that they were not due to any fault in experimental technique (Figure 15).

Fluctuations Under the Effect of Radiations. In previous publications [34, 35] two series of experiments are mentioned. In one, with 300 mg of radium sealed in glass and placed in a liquid, the sample is described
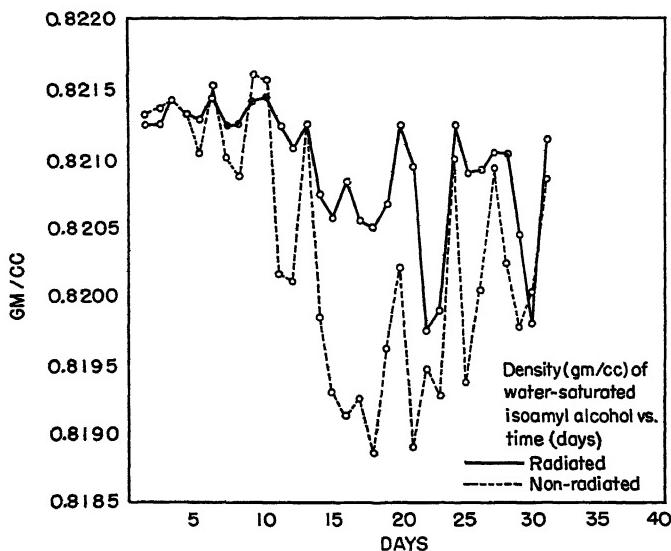
Figure 16. Fluctuations in properties of radiated liquids.

as "radiated"; the other sample, described as "non-radiated," was under the effect of the same quantity of radium placed at a distance of five feet and screened by two inches of lead. The fluctuations showed even a larger amplitude in the latter case, probably due to secondary radiation (Figure 16).

At the same time, observations were made with homogeneous liquids.* Solutions of aluminum sulfate in water showed changes in weight amounting to about 5 mg of the weight of a 25-cc pycnometer, an effect well above the limits of experimental error. Experiments recording fluctuations of the same magnitude with systems which quiet down on reaching the state of equilibrium make it quite evident that these effects

* On December 19, 1947, a letter was written to the *Phys. Rev.* by J. B. Hawkes and R. W. Astheimer, contradicting this report.[34] But these authors did not work in the presence of radiation. Secondly, they measured refractive indices and not densities.

are real. Fluctuations of the same erratic character were also recorded with ethyl alcohol, although with a smaller amplitude.

At present it appears that the amplitude of the fluctuations is correlated with the length of the chain in the organic compound in solution. Fluctuations of such magnitude support the theory of aggregates. Liquids built up according to the van der Waals' theory would not show these effects.

VI. Historical Survey and Conclusion

The science of liquids has been neglected in recent years because general interest was directed towards atomic phenomena and the solid state. X-ray analysis revived interest in liquids, but the information thus obtained is incomplete. It is therefore necessary to consider the question of molecular aggregation, which has a long history. Mendeleiev [36] stated, in 1887, that the properties of solutions are *discontinuous functions* of concentration. Extensive work on liquids was carried out by him and his school, but his conclusions have remained unknown to the outside world. Properties of liquids as a function of *temperature* are also discontinuous, this effect being shown in the third decimal of *densities*. When Antonoff's theory led to this conclusion, systematic investigation confirmed his theory as a general law. The then-current theories were at variance with this law, but Antonoff succeeded in convincing Arrhenius of the validity of this law. The latter transmitted Antonoff's manuscript to the *Z. Phys. Chem.*,[37] where it was translated by the editor, Karl Drucker, who shared the views expressed and had a similar theory of his own.[38] Following publication of his paper,[1] Antonoff received word that Geissler had observed discontinuities in benzene.[4]

Antonoff's theory resembles that of de Heen,[39] who had no experimental data to substantiate it at the time. De Heen's theory was superseded by that of van der Waals, which reigned for half a century. Finally van der Waals changed his views and recognized aggregates.[40] A good review of the literature up to 1929 can be found in a paper by Longinescu.[41] A more recent monograph by Frenkel gives the literature up to 1946.[42] This author, however, takes a one-sided view and ignores completely the extensive work on liquids by Mendeleiev and his school. The slow progress made in studies of liquid state was due to the failure of the investigators to recognize aggregates.

The theory of Duclaux [20] accidentally falls in line with that herein proposed. This author suggested, as a matter of expediency, that deviations from gas laws be attributed to associations in order to make it possible to apply the laws of "chemical mechanics." It is not customary to apply quantum mechanical treatment to liquids. Debye has emphasized

that no substantial advantages can be gained from it, and therefore Andrade, in his theory of viscosity, avoids this approach.[43] However, Born recently stated[44]: It is found that the quantum laws governing the behavior of matter in bulk are identical with classical laws, but the atomistic interpretation of the quantities appearing in the equations show great deviations from 'classical' liquid behavior.

If the current concept of molecules is replaced by that of aggregates, the kinetic theory formulated for gases works perfectly when applied to liquids. Each aggregate is an individual particle, subject to the laws of kinetic theory, as demonstrated by Antonoff. The relationships here deduced are exact because they are based on the laws of equilibrium—a fundamental concept of Gibbs. They give a definite indication as to the *number* of particles but not their nature and composition, which are better revealed by optical or x-ray methods.*

REFERENCES

1. Antonoff, G., *Phil. Mag.*, 1, 265 (1925).
2. ———, *J. Phys. Chem.*, 36, 2413 (1932).
3. Young, S., *Sci. Proc. Roy. Dublin Soc.*, 12, 374 (1910).
4. Geissler, J., *Z. Electrochem.*, 24, 101 (1918).
4a. Antonoff, G., *Réunion International de Chimie Physique, Paris*, 274 (1928).
5. ———, *J. Phys. Chem.*, 48, 80 (1944).
5a. ———, and Conan, R. J., *Science*, 109, 255 (1949); *Z. Naturforschung*, 42, 156 (1949).
6. Antonoff, G., *J. chim. phys. (Genève)*, 5, 364 (1907).
7. Freundlich, H., "Colloid and Capillary Chemistry," trans. Hatfield, pp. 86, 95, 99, New York, E. P. Dutton & Co., Inc., 1922.
8. Antonoff, G., Chanin, M., and Hecht, M., *J. Phys. Chem.*, 46, 492 (1942).
9. Adam, N. K., "Physics and Chemistry of Surfaces," p. 370, Oxford, Clarendon Press, 1938.
9a. Antonoff, G., *Rec. trav. chim.*, 68, 1011 (1949); *Experientia*, 6 (2), 52 (1950).
10. ———, *Phil. Mag.*, 1, 1263 (1926); IV, 792 (1927); *Nature*, 121, 93 (1928).
11. Dorsey, N. E., "Measurement of Surface Tension," Scientific papers of the Bureau of Standards, No. 540, p. 570, Nov. 24, 1926.
12. Antonoff, G., *J. Phys. Coll. Chem.*, 52, 969 (1948).
13. Andreas, J. M., Hauser, E. A., and Tucker, W. B., *J. Phys. Chem.*, 42, 1001 (1938).
14. Carter, E. C., and Jones, D. C., *Trans. Faraday Soc.*, 30, 11, 1027 (1934).
15. Harkins, W. D., and Ginsberg, B., *Colloid Symposium Monographs*, 6, 24 (1928); Harkins, W. D., and Humphrey, E. C., *J. Am. Chem. Soc.*, 38, 236 (1916).
16. Patrick, W. A., and Eberman, N. F., *J. Phys. Chem.*, 29, 220 (1925); Shereshefsky, J. L., *J. Am. Chem. Soc.*, 50, 2906 (1928).
17. Antonoff, G., *Ann. Physik*, 35, 84 (1939); Nellensteyn, F. J., and Roodeburg, N. M., *Chem. Weekblad*, 30, 189 (1933).
17a. von Laue, M., *Z. Krist.*, 105, 124 (1943).
18. Antonoff, G., *Phil. Mag.*, 44, 62 (1922); *Z. physik. Chem.*, 102, 388 (1922).
19. Joffé, A., International Congress of Physics, London, Oct., 1934. See also Smekal, A., *Naturwissenschaften*, 10, 799 (1922); Bragg, W. H., in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. I, p. 265, New York, Reinhold

* This work was made possible through the backing and understanding help the author received from the following scientists: F. C. Beilstein, P. P. von Weimarn, Ph.— A. Guye, Lord Rutherford, J. W. Nicholson, Sir Arthur Schuster; and in this country he owes a debt of gratitude to Wilder D. Bancroft and Karl T. Compton.

Publishing Corp., 1926; McBain, J. W., and Alexander, J., in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. III, p. 9, New York, Reinhold Publishing Corp., 1931.

20. Duclaux, J., *J. phys. (Paris),* Ser. 6, **5**, 331 (1924); Ser. 6, **8**, 336 (1927); Ser. 7, **8**, 277 (1937).
21. Antonoff, G., *Phil. Mag.,* **38**, 417 (1919).
22. McBain, J. W., *Rev. Modern Physics,* **21**, 322 (1949) [in an article by J. C. Henniker].
23. Bhagavarten, S., "Scattering of Light and the Raman Effect," p. 221, Brooklyn, N. Y., Chemical Publishing Co., Inc., 1942.
24. Antonoff, G., Hecht, M., and Chanin, M., *J. Phys. Chem.,* **45**, 791 (1941).
25. Antonoff, G., *Phil. Mag.,* **36**, 394 (1918); 1, 1132 (1926); *J. Phys. Chem.,* **36**, 2413 (1932); *Arch. Biochem.,* **6**, 199 (1945); Rhodes, F. H., and Markley, A. L., *J. Phys. Chem.,* **25**, 531 (1921).
26. Nellensteyn, F. J., World Petroleum Congress, London, July, 1933.
27. Einstein, A., in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. I, p. 323, New York, D. Van Nostrand Co., Inc., 1926. Smoluchowski, M., *Bull. Acad. Sci. Cracovie,* **218** (1916).
28. Antonoff, G., *Phil. Mag.,* **3**, 571 (1927).
29. Mark, H. F., "Physical Chemistry of High Polymeric Systems," p. 776, New York, Interscience Publishers, Inc., 1940.
30. Bragg, W. H., *Nature,* **3**, 428 (1923); Raman, C. V., *ibid.;* Oseen, C. W., *ibid.*
31. Stewart, G. W., *Phys. Rev.,* **30**, 332 (1927); **31**, 174 (1928); **35**, 726 (1930); **37**, 9 (1931).
32. Antonoff, G., "Discussion of the International Conference in Physics," Vol. **2**, "The Solid State of Matter," p. 115, London, 1934.
33. Baker, B. J., *J. Chem. Soc.,* **121**, 568 (1922).
34. Antonoff, G., *et al., Bull. Am. Phys. Soc.* (Dec. 1947); *The Science Counselor* (Duquesne University Press), **XIII**, 55 (1950).
35. ———, and Yakimac, A., *Ingenieria Quimica (Medellin),* 2 (6), 14 (1948); *Compt. rendu,* (May 22, 1950).
36. Mendeleiev, D., *Z. physik. Chem.,* **1**, 283 (1887).
37. Antonoff, G., *Z. physik. Chem.,* **112**, 461 (1924).
38. Drucker, K., *Z. physik. Chem.,* **68**, 616 (1909).
39. de Heen: see Mathias, "Le Point Critique des Corps Purs," Paris, 1904.
40. van der Waals, *Proc. Koninkl. Akad. Wetenschappen Amsterdam,* **13**, 1253 (1911).
41. Longinescu, G. G., *Chem. Revs.,* **6**, 381 (1929).
42. Frenkel, J., "Kinetic Theory of Liquids," Oxford, Clarendon Press, 1946.
43. Andrade, E. N. da C., *Phil. Mag.,* **17**, 698 (1934); *Nature,* **125**, 309, 580 (1930).
44. Born, M., and Green, H. S., *Nature,* **159**, 738 (1947).
45. Turner, W. E. S., "Molecular Association," London, Longmans Green, 1915; also his paper in Vol. **I** of this series (1926).
46. Daniels, F., Mathews, J. H., Williams, J. W., *et al.,* "Experimental Physical Chemistry," pp. 445-6, New York, McGraw-Hill Book Co., Inc., 1929.

FERROMAGNETISM IN THE COLLOIDAL PARTICLE

L. J. E. Hofer

Central Experiment Station, Bureau of Mines, Pittsburgh, Pa.

The Physical Basis of Magnetism

Very few direct experimental data on the nature of ferromagnetism in particles of colloidal size are available. Nevertheless, this problem vitally affects many important subjects such as the chemical processes involving iron, cobalt, and nickel as catalysts; the fluidization properties of ferromagnetic particles; and suspensions of ferromagnetic particles which are used in the magnetic clutch. The basic aspects of the relationships between particle size and ferromagnetic properties will be outlined here.

All matter can be conveniently classified as diamagnetic, paramagnetic, or ferromagnetic.[1-8, 19, 27-29, 32, 33] Paramagnetism is caused by a permanent atomic or molecular moment resulting from the spin of a single unpaired electron or from the resultant spin of an uncompleted shell of electrons. That paramagnetism is caused mainly by unbalanced electronic spin has been convincingly demonstrated by gyromagnetic experiments.[2] The magnetic moment associated with nuclear spin and with the orbital motion of electrons is, in general, very minor compared with that produced by electronic spin itself. Ferromagnetism is a special case of paramagnetism. It occurs if the parallel alignment of the magnetic dipoles is thermodynamically more stable than the disordered arrangement normally exhibited by paramagnetic substances. By virtue of its ordered arrangement, the ferromagnetic state has a lower entropy than the disordered paramagnetic state and will occur only when favored by low temperatures. As a result, every ferromagnetic material can exist in both its characteristic ferromagnetic condition at low temperatures and in a paramagnetic condition at high temperatures. The transition temperature is known as the "ferromagnetic Curie point." It should be pointed out that the orientation of the magnetic dipoles is relatively independent

of the orientation of the crystal structure, although there are certain crystal axes along which magnetization is easiest. The forces in ferromagnetics which make the oriented condition more stable than the unoriented condition of the magnetic dipoles will be discussed later. In general, ferromagnetism can be considered as a cooperative phenomenon of a group of atoms or molecules and is not the result of the individual action of atoms or molecules. Ferromagnetism is thus never exhibited by liquids or gases, but only by solids or by solids dispersed in fluid media and should, therefore, be of particular interest to colloid science.

All substances, whether or not they contain permanent magnetic dipoles, exhibit the phenomenon of diamagnetism. The diamagnetism is simply superimposed on the paramagnetism or ferromagnetism, if either one is also present. For convenience in classification, only those materials which are neither ferromagnetic nor paramagnetic are designated as diamagnetic. The phenomena of diamagnetism are all based on the induction of a magnetic dipole by means of an external magnetic field.

The following may be noted as practical characteristics of diamagnetism, paramagnetism, and ferromagnetism: Diamagnetics tend to be repelled by a magnetic field, i.e., they tend to move in a direction away from the greater field densities; paramagnetics are attracted by a magnetic field, i.e., they move in the direction of greatest field density. (An exception to the above is found in molecular diamagnetics where, under unusual circumstances, a molecule having no permanent magnetic dipole may act as a feeble paramagnetic.) In general, the magnitude of the force of repulsion on a diamagnetic is much weaker than that of the force of attraction on a paramagnetic when these are placed in comparable magnetic fields.

The force of attraction on a ferromagnetic is of an entirely different order of magnitude when compared with that of a paramagnetic. Only if a body is ferromagnetic will it respond perceptibly to the fields ordinarily available with small portable permanent magnets. Probably the most characteristic feature of ferromagnetics is the Curie point at which ferromagnetism changes to paramagnetism. The Curie points of the ferromagnetic elements are: α-iron, 760°C; cobalt, 1120°C; nickel, 360°C; and gadolinium, −257°C (16°K). Ferromagnetism is not common in nature, although it is probably more common than is generally realized. Many of the oxides, carbides, nitrides, hydrides, and alloys of these ferromagnetic elements are also ferromagnetic. Of particular interest are the ferromagnetic alloys and compounds of materials such as manganese and chromium which are normally nonferromagnetic. These ferromagnetics include some of the nitrides, phosphides, arsenides, stibides, silicides, stannides, hydrides and borides of manganese and also some of the oxides, tellurides, and platinum alloys of chromium. The Heusler–type

alloys—$AlMnCu_2$, $MnSnCu_2$, $SnMn_3Cu_6$, Ag_5MnAl, and others—are ferromagnetic and of great theoretical interest.[16, 17]

Ferromagnetism

For further development of our subject, it will be necessary to define magnetic moment, intensity of magnetization, and specific magnetization. Consider a permanent magnet of length l and pole strength m placed in a field H and making an angle ϕ with the direction of the field. Each pole will experience a force of

$$F = mH \tag{1}$$

and the two forces will form a couple whose moment is

$$L = lmH \sin \phi \tag{2}$$

Since L, H, and ϕ can be independently and experimentally determined, the quantity ml is exactly defined even though, independently, l and m are incapable of exact definition. This quantity, ml, is called M, the magnetic moment. It can be shown that the force, F, experienced by a magnet of moment, M, is simply

$$F_x = M \frac{dH}{dx} \tag{3}$$

The intensity of magnetization, I, of a magnet is defined as M/V, where V is the volume of the magnet. Similarly, the specific magnetization, σ, of the magnet is M/W, where W is the mass of the magnet.

Since the magnetic order of the ferromagnetic material is destroyed at the Curie point, θ, the energy involved must be of the order $k\theta$, where k is the Boltzmann constant. This energy may be evaluated in terms of a field H and the magnetic moment, $\beta\mu$, of the individual magnetic dipole, where β is the magnetic moment of the spinning electron (Bohr magneton per mole), and μ is the number of Bohr magnetons of the specific ferromagnetic. Thus,

$$k\theta \approx H\beta\mu \tag{4}$$

For metallic iron, $\theta = 1040°K$ and $\mu = 2.22$.

$$H \approx \frac{1040 \times 8.314 \times 10^7}{5563 \times 2.22} \approx 7.03 \times 10^6 \text{ gauss} \tag{5}$$

This is an exceedingly strong field which has not been achieved experimentally to date. The ordering of the individual magnetic dipoles cannot, therefore, be attributed to any external field; nor can it be due to the magnetic interactions of the individual dipoles themselves, for calculations show that magnetic interaction is barely sufficient to account for a Curie point of about 1°K.

Weiss [27, 32] circumvented this dilemma by assuming the existence of an "internal field" of the order of 10^4 times the magnetic interaction of the elemental magnetic dipoles. With this assumption, Weiss was able to describe mathematically, and with a fair degree of accuracy, many phenomena of ferromagnetism. As a necessary corollary to the "internal field," Weiss was forced to conclude that ferromagnetic material was composed of many tiny volumes, each completely magnetized to saturation in a direction of easy magnetization. In the unmagnetized state below the Curie point, these tiny volumes are so arranged that their magnetic effects mutually cancel. When a gradually-increasing magnetic
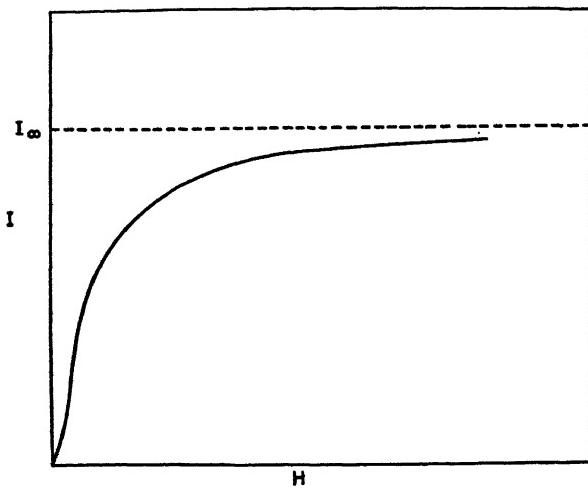
Figure 1. Schematic plot of intensity of magnetization, I, as a function of field, H, for the initial magnetization of a ferromagnetic.

field is applied to an "unmagnetized" ferromagnetic, the resulting magnetization is observed to proceed in three successive stages as follows: (1) Those domains which have a component of magnetization in the direction of the applied field tend to grow at the expense of other domains; (2) those domains having a negative component in the direction of the applied field tend to reverse their magnetization suddenly, an irreversible process which results in tiny discontinuities (Barkhausen effect) of the magnetization curve; (3) the magnetic orientation of the domains which, by the end of process (2), lay in one of the directions of easy magnetization, are rotated with increasing field strength to coincide with the direction of the applied field. This latter process requires very high field strengths. The directions of easy magnetization are: [100] for α=Fe, [111] for nickel, [001] for cobalt below 270°C, and [100] for cobalt above 270°C.[5] The schematic magnetization curve is shown in Figure 1.

The maximum magnetization for infinite field at any given temperature is designated by I_∞. I_∞ is a function of temperature only, varying from zero at the Curie point to $I_{\infty,0}$ at the absolute zero, where it is a maximum. $I_{\infty,0}$ is a characteristic constant of each ferromagnetic. When $I_\infty/I_{\infty,0}$ is plotted against the corresponding T/θ, the resultant curves for most ferromagnetics very nearly coincide (Figure 2).

Heisenberg [32] showed that the "molecular field" of Weiss is really an exchange energy of electrostatic origin. This exchange energy is best
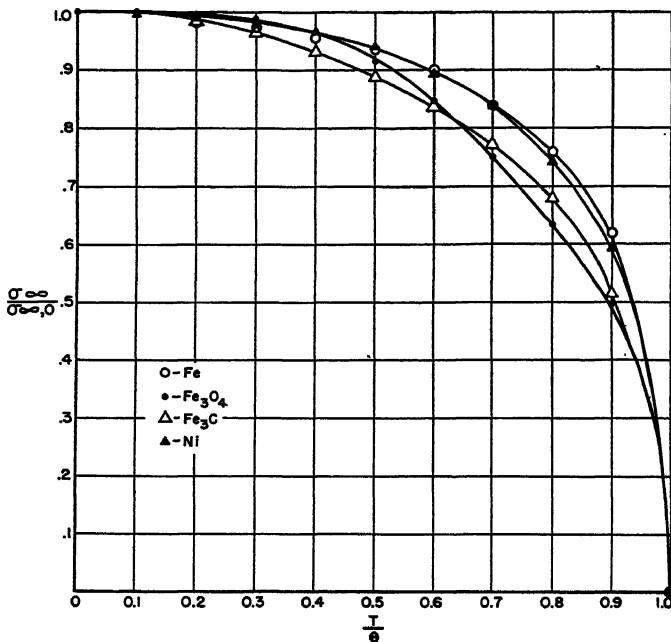
Figure 2. Reduced thermomagnetic curve of Fe, Fe_3O_4, Fe_3C, and Ni.

measured in terms of J, the exchange integral, or T_c, a mathematically derived Curie temperature. Two cases of the Heisenberg theory of ferromagnetism are readily solvable. For the first case, S, the spin quantum number, is assumed equal to $1/2$, and the energy states of a crystal belonging to a particular spin are assumed to be distributed as a continuum according to a Gaussian distribution. This case leads to an expression for the Curie temperature as follows:

$$T_c = 2JZS(S+1)/3k \tag{6}$$

where Z is the number of neighbors of each atom. In the second case, S is allowed to be arbitrary, but the less realistic assumption that all the

crystalline states of the same spin have the same energy must be made. These assumptions lead to an expression for the Curie point of

$$T_c = 2J/k(1 - \sqrt{1 - 8/Z}) \qquad (7)$$

Above the Curie point, Heisenberg's considerations lead to the following equation for magnetic susceptibility:

$$\chi = 4N\beta^2 S(S + 1)/3k(T - T_c) \qquad (8)$$

where N is the number of molecules per unit volume, β is the Bohr magneton (0.9174×10^{-20} e.m.u.); the other symbols have been previously defined. It will be seen that equation (8) has the form $\chi = C/(T + \Delta)$ which Curie found to describe the behavior of most paramagnetics. (Δ which equals $-T_c$ has no significance for ferromagnetism when positive. When Δ is negative, it is known as the paramagnetic Curie point. For most ferromagnetics, the paramagnetic Curie point is some 10 to 20°C higher than the ferromagnetic Curie point. A negative Δ is not necessarily a criterion of ferromagnetism.)

Equation (6) shows that, with the assumption of a Gaussian energy distribution for the crystalline states, the necessary condition for ferromagnetism is that the exchange integral, J, be positive. Even when J is positive, a real Curie temperature may not exist if a non-Gaussian distribution of energy levels is assumed. Equation (7) shows that a real T_c is achieved only if $Z \geq 8$. Heisenberg points out that the common ferromagnetics, iron, cobalt, and nickel, are either body-centered cubic ($Z = 8$), face-centered cubic ($Z = 12$), or close-packed hexagonal ($Z = 12$). Thus, there may be other criteria for ferromagnetism in addition to a positive J.

Slater,[31] in expanding on Heisenberg's work, has shown that two conditions must be met in order to have a sufficiently large positive J for ferromagnetism: (1) There must be an uncompleted quantum shell of high-azimuthal quantum number (d or f); and (2) the radius of this shell must be small compared with the interatomic distance of the nearest atoms of the same element. The ratio, v, of the interatomic distance to the radius of the uncompleted quantum shell has been calculated by Slater, and the results are given in Table 1. From these data and equation (6), a semi-quantitative relation between J and v can be determined (Figure 3). This graph suggests that elements such as manganese and chromium can become ferromagnetic if the interatomic distance is increased by some means so that the ratio v falls in the ferromagnetic range. This is precisely what seems to occur in the case of manganese nitride, carbide, and hydride, which are well-known interstitial compounds and are definitely ferromagnetic. In these compounds, the lattice parameters are considerably greater than would be expected if the manganese atoms were separated by the normal Mn–Mn distance found in manganese

metal. Presumably, the ferromagnetism of the arsenides, phosphides, stibides, bismuthides, sulfides, selenides, tellurides, silicides, and stannides is attributed to a similar relationship. Similarly, certain chromium tellurides, hydrides, arsenides, and oxides are ferromagnetic. Ferromagnetism is not as common among chromium compounds as among manganese compounds because of the greater increase in interatomic distance
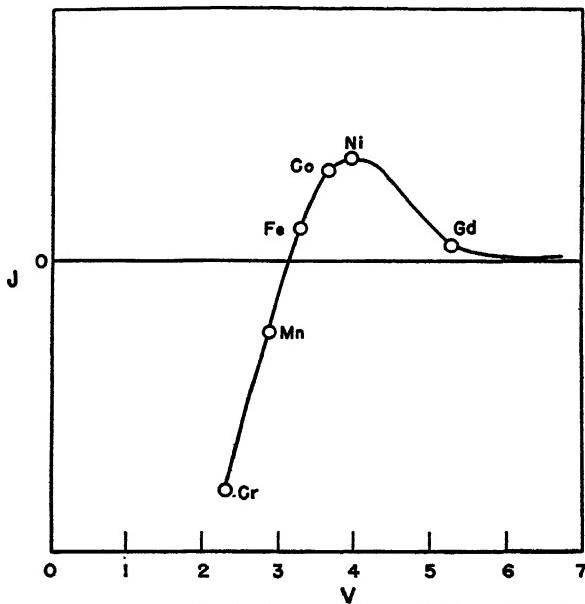
Courtesy Bureau of Mines, U.S. Dept. of Interior

Figure 3. Schematic plot of the exchange integral, *J*, as a function of the ratio of interatomic distance to quantum shell radius.

necessary to bring the compound into the ferromagnetic range. An interesting case of this same phenomenon is presented by the Heusler alloys.[16, 17, 24-26] In these alloys, under proper annealing conditions, a super-lattice is believed to be formed in which the Mn–Mn distance is in the ferromagnetic range. Generally, the term "Heusler alloys" includes Mn–Al–Cu and the Mn–Cu–Sn alloys. In these ternary systems, the compositions $MnAlCu_2$, $MnSnCu_2$, and $SnMn_3Cu_8$ give maximum specific magnetizations. For the first composition, x-ray studies have definitely proved the existence of a super-lattice. Arsenic, antimony, bismuth, and boron can replace aluminum in these alloys without materially affecting the ferromagnetism. A closely related compound, Ag_5MnAl, is also ferromagnetic.

TABLE 1. THE RATIO OF INTERATOMIC DISTANCE TO
QUANTUM SHELL RADIUS FOR
FERROMAGNETICS AND NEAR-FERROMAGNETICS

Ti	Cr	Mn	Fe	Co	Ni	Pd	Pt	Ce	Yb
2.24	2.36	2.94	3.26	3.64	3.96	2.82	2.46	3.20	5.28

On the basis of purely experimental data, Dehlinger [6] has proposed to determine the course of the curve for J versus interatomic distance. Dehlinger's J is, however, a composite involving not only the interactions of nearest neighbors but also the interactions of neighbors farther removed. Dehlinger has been quite successful in applying his theory to the study of alloys.

Slater [30] and others have been successful in applying the band theory of metals to the properties of ferromagnetics. From this viewpoint, the electrons are considered free, with minor restrictions imposed on that freedom by the electrostatic fields of the metal ions. According to this theory, the $3d$ and $4s$ shells of iron, cobalt, and nickel are incompletely filled, and the manner of distribution of the electrons between these shells determines the magnetic moments. This theory has proved useful in determining how I_∞ varies with composition in certain ferromagnetic alloys.

Colloidal Phenomena and Ferromagnetism

As has been already mentioned, a ferromagnetic, even in its demagnetized state, consists of domains which are uniformly magnetized to saturation. Although the domain theory has been very useful, it has been difficult to obtain much information concerning the shapes of the domains by either a theoretical or experimental approach. Elmore and McKeehan [11] have studied the patterns produced by applying a colloidal suspension of magnetite to carefully polished ferromagnetics. Later, it was shown that this technique could yield more characteristic results if the ferromagnetic was electrolytically polished. Very recently, Williams [35, 36] and his co-workers have succeeded in interpreting the powder patterns in some detail. There are three forms of energy which determine the size and shape of the magnetic domain.[18] These energies are the surface energy of the domain boundaries, F_w, the magnetic field energy of the configuration, F_m, and the anistropy energy of spin orientation, F_a. That domain geometry is assumed which gives the lowest total energy for all these different forms. The energy of the Bloch wall or domain boundary can be written

$$F_w = w/A \qquad (9)$$

where A is the surface area of the domain and w is the energy associated

with the surface per unit area. This energy has been calculated to be about 1 to 5 ergs/cm². The magnetic energy, F_m, is given by the formula

$$F_m = -1/2 \int (H \cdot I) dV \tag{10}$$

where I is the intensity of magnetization, and H is the magnetic field produced by the magnetization. H, more specifically, is NI, the product of the demagnetizing factor, N, and the intensity of magnetization. Therefore,

$$F_m = NI^2V/2 \tag{11}$$

The anisotropy energy is

$$F_a = \rho_a V_a \tag{12}$$

where ρ_a is the anisotropy energy, per unit volume, of domains oriented in directions away from an axis of easy magnetization, and V_a is the total volume of the domains in the specimen not oriented near an easy direction. ρ_a is of the order 10^5 to 10^6 ergs/cm³. From these relationships it

Figure 4. Domain configurations in small particles. (After Kittel.[18])
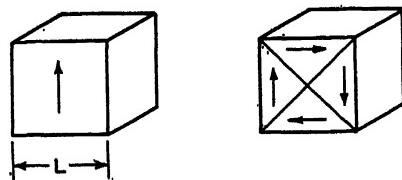
should be obvious that as the particle size decreases, F_w, the domain boundary energy, becomes large as compared with volume energies, F_a and F_m. Thus, a point is eventually reached where a structure containing Bloch walls is metastable with respect to one which does not have them. In other words, each particle will be a single domain and will act as a permanent magnet. Kittel's [18] calculation for a tiny ferromagnetic cube is very instructive. Consider particles in the form of cubes of edge length L, one containing a single domain and one containing four domains as shown in Figure 4. In the case of the particle containing a single domain, the energy is entirely magnetic and is given by

$$F_m = (2\pi/3)I^2L^3 \tag{13}$$

In the case of the particle containing four domains, the flux circuit is closed internally so that $F_m = 0$. The wall energy is

$$F_w = 2\sqrt{2}\,w L^2 \tag{14}$$

The anisotropy energy is approximately

$$F_a = \rho_a L^3/2 \tag{15}$$

The critical case is the one where the energy of the particle containing four domains equals the energy of the particle containing a single domain. From this relationship, the critical diameter can be calculated

$$L_c = \frac{2\sqrt{2}w}{(2\pi/3)I^2 - \rho_a/2} \approx 1.5 \times 10^{-6}\,\text{cm} \tag{16}$$

Direct experimental evidence that small particles of magnetite behave as permanent magnets has been obtained by Elmore.[9, 12, 13] In order to interpret his results, it was necessary to postulate that each colloidal particle behaves as a molecule of a paramagnetic gas. The experimental results of Montgomery [21] and Heaps [14] essentially support Elmore.

Other features of colloidal particles of a single domain are low initial permeability and high coercive force. This is related to the fact that magnetization changes cannot occur through movements of the domain boundaries but only by spin rotation. Spin rotation is opposed by anisotropy forces which are strong compared with the forces opposing Bloch wall movement. There are many examples in the literature where small particles or films exhibit unusually high coercive forces and low permeabilities.

The magnetic domains, in general, do not coincide in extent, shape, or size with the crystallites of a ferromagnetic. If the crystallites are smaller than the critical size for a magnetic domain, the domain will be bounded by the crystallite boundaries if they are thick enough to isolate the grain magnetically from the next grain. As an upper limit to this isolating thickness of grain boundary, the distance (5 to 10Å) at which J becomes relatively small must be chosen.

Ferromagnetism is the result of exchange forces between ions in a crystal and is never observed in the liquid or gas phase or in true solutions, probably because of the greater separation of ions in these phases. Under these conditions, no matter how large the permanent atomic magnetic moment is, only paramagnetism will be observed. We may conclude, therefore, that, in passing from large magnetically isolated particles to magnetically isolated atoms by decrease in particle size, a transition from ferromagnetism to paramagnetism must occur. Haul and Schoon,[15] Winkel and Haul,[37] and Beischer and Winkel,[4] by following the magnetic susceptibility, have studied this transition in tiny spherical particles of $\gamma - Fe_2O_3$ produced by the oxidation of carbonyl iron. They found the transition at 40Å. König [20] studied the transition by means of the Faraday effect on light transmitted through thin evaporated films of iron and found that below a thickness of 10 to 12Å the iron was not ferromagnetic. König attempts to correlate the value of this transition size with lattice parameter, but this seems questionable. These relationships are illustrated schematically in Table 2. Material composed of

particles which are slightly smaller than the critical size for the ferro-magnetism–paramagnetism transition probably still has a high paramagnetic Curie point, but, as the particle size approaches that of a single atom, the paramagnetic Curie point approaches zero, and the material is magnetically dilute.[7, 8]

TABLE 2. EFFECT OF PARTICLE SIZE ON THE
MAGNETIC PROPERTIES OF MAGNETICALLY
ISOLATED FERROMAGNETIC PARTICLES

Type of Magnetism	Polydomain Ferro-magnetic	Monodomain Ferro-magnetic	Nondilute Para-magnetic	Dilute Para-magnetic
Permeability of particle	High	Low	Exceedingly low	Exceedingly low
External field of particle	Low	Exceedingly high	Exceedingly low	Exceedingly low
Paramagnetic Curie point, $\triangle$	Near θ	Near θ	$>0°K$	$\leqq 0°K$

(Transitional Particle Size (cm): 10^{-6} to 10^{-5} ; 5×10^{-7} to 10^{-7} ; 5×10^{-5})

Transitional Particle Size (cm)

REFERENCES

1. Bates, L. F., "Modern Magnetism," Cambridge, Cambridge Univ. Press, 1939.
2. Barnett, S. J., *Rev. Modern Phys.*, **7**, 145 (1935).
3. Becker, R., and Döring, W., "Ferromagnetismus," Berlin, Julius Springer, 1939.
4. Beischer, D., and Winkel, A., *Naturwissenschaften*, **25**, 420 (1937).
5. Bozorth, R. M., *Rev. Modern Phys.*, **19**, 29-86 (1947).
6. Dehlinger, U., "Chemische Physik der Metalle und Legierungen," Leipzig, Akademische Verlagsgesellschaft m.b.H., 1939.
7. Eischens, R. P., and Selwood, P. W., *J. Am. Chem. Soc.*, **69**, 1590 (1947).
8. ———, and ———, *J. Am. Chem. Soc.*, **69**, 2698 (1947).
9. Elmore, W. C., *Phys. Rev.*, **60**, 593 (1941).
10. ———, *Phys. Rev.*, **51**, 982 (1937).
11. ———, and McKeehan, L. W., *Trans. Am. Inst. Mining Met. Engrs.*, **120**, 236 (1936).
12. ———, *Phys. Rev.*, **54**, 1092-5 (1938).
13. ———, *Phys. Rev.*, **54**, 309-10 (1938).
14. Heaps, C. W., *Phys. Rev.*, **57**, 528 (1940).
15. Haul, R., and Schoon, T., *Z. Elektrochem.*, **45**, 663 (1939).
16. Heusler, O., *Ann. Physik*, **19**, 155 (1934).
17. Heusler, Fr., *Verhandl. deut. phys. Ges.*, **5**, 219 (1903).
18. Kittel, C., *Phys. Rev.*, **70**, 965 (1946).
19. Klemn, W., "Magnetochemie," Leipzig, Akademische Verlagsgesellschaft m.b.H., 1936.
20. König, H., *Naturwissenschaften*, **33**, 71-5 (Aug. 15, 1946).
21. Montgomery, C. G., *Phys. Rev.*, **39**, 162 (1931).
22. Néel, L., *Compt. rend.*, **224**, 1488-9 (May 28, 1917).
23. ———, *Ann. Phys.*, **8**, 237-308 (1937).
24. Persson, E., *Z. Physik*, **57**, 115 (1929).
25. Potter, H. H., *Proc. Phys. Soc. London*, **41**, 135 (1929).
26. ———, *Phil. Mag.*, (7), **12**, 255 (1931).
27. Seitz, F., "The Modern Theory of Solids," New York, McGraw-Hill Book Co., Inc., 1940.

28. Selwood, P. W., "Magnetochemistry," New York, Interscience Publishers, Inc., 1943.
29. ———, *Chem. Revs.*, **38**, 41-82 (1946).
30. Slater, J. C., *J. Applied Phys.*, **8**, 385 (1937).
31. ———, *Phys. Rev.*, **52**, 198 (1937).
32. Van Vleck, J. H., "The Theory of Electric and Magnetic Susceptibilities," Oxford, Clarendon Press, 1932.
33. ———, *Rev. Modern Phys.*, **17**, 27 (1945).
34. Wannier, G. H., *Rev. Modern Phys.*, **17**, 50 (1945).
35. Williams, H. J., Bozorth, R. M., and Shockley, W., *Phys. Rev.*, **75**, 155-78 (1949).
36. ———, and Shockley, W., *Phys. Rev.*, **75**, 178 (1949).
37. Winkel, A., and Haul, R., *Z. Elektrochem.*, **44**, 823 (1938).

LUMINESCENT SOLIDS (PHOSPHORS)

H. W. Leverenz

RCA Labs. Div., Radio Corp. of America, Princeton, N. J.

IT IS A CURIOUS FACT that man synthesized and used luminescent solids long before detecting luminescence from the natural luminescent solids which have existed on this earth since its genesis. About 350 years ago, the alchemist, Casciarolo, chanced to heat some native barium sulfate with charcoal and noticed that after exposure to daylight the cooled impure sulfide product glowed feebly in the dark. This mysterious solid, which was called a *phosphor* or phosphoros (Greek, "light bearer"), antedated the discovery of the chemical element phosphorus by about 70 years. Because the similarity of terms has caused some confusion, it should be noted that the luminescence of a phosphor is a solid-phase *physical* (electronic) action which occurs throughout the mass and which may continue indefinitely when the phosphor is in a vacuum; whereas the luminescence of the element misnamed phosphorus results from a gas-phase *chemical* action which occurs only at the surface and which ceases when the phosphorus is consumed or placed in a vacuum.

At the time phosphors were first prepared, daylight was the only known means for their excitation, and so luminescence radiation was observable only when it was excitable by daylight and persisted long enough for the phosphor to be taken into a dark place. With the development of modern electronic and radioactive sources of invisible emanations, phosphors have been used not only for instantaneous detection of ultraviolet, x-rays, cathode rays, alpha particles, etc., but also as a means of putting these invisible forms of energy to work in television, radar, electron microscopes, "fluorescent" lamps, infrared sniperscopes, x-ray fluoroscopes, and self-luminous dial markings. This article briefly outlines the present status of man-made phosphors, whose variety and capabilities greatly exceed those of the known natural luminescent solids. More detailed information may be obtained from the references at the end of this article.

Before proceeding, it is worth emphasizing that the technology of phosphors, like that of other structure- and impurity-sensitive, electronically-

125

active solids (e.g., photoconductors, semiconductors, ferroelectric and ferro-magnetic materials), is in the process of growing from the status of an art to that of a science. Our imperfect understanding and control of electronically active solids is caused largely by the practical impossibility of completely segregating (purifying) and rectilinearly arranging (crystallizing) the enormous number of atoms ($10^{23}/cm^3$) in *real* crystals, so that real crystals are always impure and imperfect to some degree. It is known that the unavoidable imperfections (including impurities) in real crystals sometimes increase and sometimes decrease electronic activities such as excitation, internal ionization, radiative and nonradiative transitions, electron mobility, and trapping. Very often, certain crystal imperfections are incorporated deliberately to promote or suppress a given electronic activity. In the case of phosphors, the highly purified *host crystal* usually has a very small proportion of *activator* impurity which is either added initially or is induced to increase luminescence efficiency. The efficiency increase is generally accomplished by decreasing the proportion of excitation energy dissipated in the crystal as heat.

Luminescence

According to the quantum interpretation of radiation phenomena, all radiant energy is absorbed and emitted as photons whose energies E are quantized according to

$$E = h\nu = hc/\lambda \text{ erg (dyne} - \text{cm)}$$

where $h = 6.624 \times 10^{-27}$ erg sec $= 4.14 \times 10^{-15}$ ev sec

          ν = the oscillation frequency of the radiation, in cycles per second ($= \sec^{-1}$)

          $c = 2.9978 \times 10^{10}$ cm sec^{-1}

          λ = the diffraction wavelength of the radiation, in centimeters.

The letters "ev" denote an *electron volt* which is a unit of energy equal to the energy acquired by an electron falling through a potential difference of one volt [1 ev $= 1.60 \times 10^{-12}$ erg $= 3.82 \times 10^{-20}$ cal $= 23.05$ kcal mole^{-1} (when each simple molecule in a mole has 1 ev)].

When a simple *pigment* absorbs radiant energy, such as visible or ultraviolet radiation, the primary photon energy is converted (degraded) into heat which *diffuses* through the material. The heated material then emits radiation, primarily in the infrared, which is called *thermal radiation*. Thermal radiation may equal but cannot exceed the thermal emissive power P_ν (at frequency ν) of a perfect black body at the same absolute temperature T. This thermal radiation for a black body follows Planck's radiation law

$$P_\nu = 8\pi h\nu^3 c^{-3}(\epsilon^{h\nu/kT} - 1)^{-1} \text{ erg cm}^{-2} \text{ sec}^{-1}$$

where $k = 1.38 \times 10^{-16}$ erg deg^{-1} $= 8.62 \times 10^{-5}$ ev deg^{-1}

The quality and quantity of thermal radiation from solids (crystals) is little

dependent on their natures and is strongly dependent on their temperatures (Figure 1).

When a suitable *luminescent solid* (*phosphor*) absorbs, for example, ultraviolet radiation, an appreciable part of the absorbed energy is temporarily *localized* as excitation of certain atoms or small groups of atoms which then emit characteristic radiation in *excess* of the thermal radiation from the solid. *Luminescence*, then, is a physical process whereby matter generates nonthermal radiation which is characteristic of the given luminescent material (Figure 1). When the temperature of a solid is increased, the intensity of any luminescence radiation generally decreases, whereas the intensity of thermal radiation always increases.
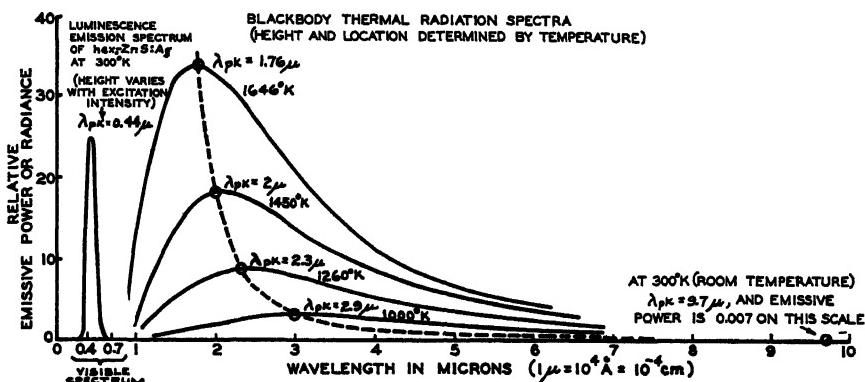
Figure 1. Comparative spectral distributions of the emissions from a typical phosphor (narrow band at left) and a black-body thermal emitter (broad bands at right).

If *isolated* atoms of, say, gaseous manganese are excited by primary photons or charged material particles (electrons, ions), the excited atoms can dispose of their excess energy only by emitting luminescence radiation. This radiation appears as sharp spectral *lines* which are characteristic of atomic manganese. The sharpness of the lines is indicative of the discreteness of the energy levels of manganese in the ground and excited state. It is only when excited free atoms collide with other atoms, or with the walls of the container, that the excited atoms have an opportunity to dispose of their excess energy as heat.

The situation is quite different, however, when the foregoing rarefied manganese gas is cooled and condenses to form first a liquid and then a solid. The solid manganese may absorb primary energy as before, but it converts this energy into heat rather than luminescence radiation (neglecting x-ray and gamma-ray radiation). In solid manganese, the outermost electrons of the atoms no longer belong to particular atoms, but are shared by all the identical atoms in the crystal. This sharing, coupled with the operation of the Pauli exclusion principle, broadens the discrete energy levels of the free

manganese atoms into *bands* in the solid. The topmost occupied energy band of solid manganese is only partially filled with very loosely bound electrons, and so absorbed energy increases slightly the energies of these nearly free electrons which then agitate the atoms and increase the heat content of the solid without producing appreciable luminescence emission.

An *approximation* to the independently-luminescing free-manganese atoms may be obtained by incorporating small proportions of manganese as a (combined) impurity in certain solids. As examples, if one out of every three-hundred zinc atoms in crystalline zinc orthosilicate or zinc sulfide is replaced by a manganese atom, then the resultant phosphor crystals exhibit luminescence which is definitely associated with the manganese impurities. The separated manganese impurity atoms (ions to some degree) and their perturbed host-crystal neighbors (ligands) function as independent luminescence-active *centers* called activator centers. The sizes of centers in various phosphors range from atomic to colloidal dimensions.

A center may be visualized as a partially disordered region wherein there is a contest between (1) the ambient host-crystal atoms which strive to maintain their normal crystalline arrangements and spacings, and (2) the central impurity atom which strives to rearrange the host-crystal atoms about it in configurations and spacings appropriate for its size, effective charge, and directional-bonding characteristics. On increasing the proportion of activator centers in phosphors, the efficiency of the luminescence process generally increases until the centers become so numerous that they overlap (interact) unduly with each other.

With respect to terminology, distinctive prefixes are used to denote luminescence produced by primary particles which differ in the manner in which they excite phosphors, e.g., visible and near-visible photons excite *photoluminescence*, x-ray and gamma-ray photons excite *roentgenoluminescence*, electrons excite *cathodoluminescence*, and ions, such as alpha particles, excite *ionoluminescence*. The energies of these primary excitant particles generally exceed 2 electron volts (ev), there being no known upper limit (at least to 10^8 ev) to the energy of a primary particle capable of exciting phosphors.

The energies of the photons emitted during *conventional luminescence* range from about 1 ev (1.24×10^{-4} cm = 12,400Å) to 10 ev (0.124×10^{-4} cm = 1240Å), while the energies of emitted x-ray and gamma-ray luminescence photons extend well beyond this range up to about 10^7 ev. This article is concerned chiefly with the conventional visible and near-visible luminescence emissions of solids, which generally involve transitions of the outer or valence electrons comprising the "skins" of atoms. These exposed electrons are sensitive to changes in the kind, number, spacing, and arrangement of neighboring atoms in a solid, and the luminescence of solids is thus a sensitive indicator of changes in composition, impurities or other imperfections, and crystal structure. Also, the valence electrons of an atom in a solid experience

most strongly the jostling caused by thermal motion of atoms, so that increasing the temperature of the solid generally perturbs the luminescence process and increases the probability of conversion of high-energy excitation quanta into lower-energy thermal quanta (phonons).

With respect to the duration of luminescence, *fluorescence* indicates a normal, unconstrained, spontaneous radiative return from the excited state (as in isolated nonmetastable atoms or ions), whereas *phosphorescence* indicates an abnormally long delay between excitation and emission, using *isolated* atoms or ions as standards of normal behavior. Fluorescence, then, is a limiting case of phosphorescence and corresponds to a natural excited-state lifetime, τ_F, of about 10^{-8} or 10^{-9} second for the nonmetastable excited states of isolated atoms or ions undergoing the optical transitions which occur in conventional luminescence. The natural lifetimes of isolated atoms are determined chiefly by oscillator damping (classically, $\tau_F \propto \nu^{-3}\rho_d^{-2}$ for radiation of photons of frequency ν from a dipole with moment ρ_d), and the width of a fluorescence emission *line* is determined by the indeterminacy of the excited-state energy level, ΔE^*, such that $\Delta E^* \geq h/2\pi\tau_F$, where $h = 6.62 \times 10^{-27}$ erg sec. This line width is only about 10^{-7} ev for conventional fluorescence with $\tau_F \approx 10^{-8}$ second, whereas many phosphors have emission bands nearly 1-ev wide and their emissions persist for seconds or days (i.e., the band widths and persistences of phosphors are often unrelated). Most phosphors exhibit predominantly an abnormally delayed emission, which is called phosphorescence. Here, the abnormal delay may be caused by (1) the strong perturbing (constraining) influence of neighboring atoms on excited centers in solids, and/or (2) internal ionization and trapping. *Internal ionization* is the ejection of an electron from an excited atom or center, without the electron's leaving the solid. The vagrant excited electron may become trapped, particularly near imperfections in the crystal, and remain trapped for an indefinite time before being released by heat or other energy so that it can again wander to make a radiative recombination with an ionized center.

In general, phosphors begin to emit luminescence photons within 10^{-8} second after onset of excitation, but very often much of the excitation energy is stored in the form of prolonged (constrained) excited states or trapped excited electrons so that photon emission is extended for intervals ranging from a few microseconds to a few years after cessation of excitation, the duration of phosphorescence depending on the nature of the phosphor and its conditions of excitation and operation. When a large proportion of the excitation energy is stored, the curve of luminescence output vs. time exhibits a detectable *growth* (Figure 2) until equilibrium is established, i.e., until the rate of filling and emptying of excited states and traps has stabilized in the excited volume of the solid. After cessation of excitation, there is a *decay* of luminescence output vs. time as the stored excitation energy is

released. When there is simple excitation, without internal ionization and trapping, the rate of emission of luminescence photons, L, decreases exponentially with time, t, according to $L = L_0 \epsilon^{-\alpha t}$, where L_0 is the luminescence output at cessation of excitation and the decay constant α ($= \tau^{-1}$) ranges from about 10^7 to 1 sec^{-1}, depending on the composition and structure of the phosphor. It is to be expected that τ will be larger and, hence, α will be smaller the more constraint an excited emitting atom or center experiences from its neighbors in the crystal, but the emitter itself is a primary factor in determining τ. In this case, the spontaneous *exponential decay* is little
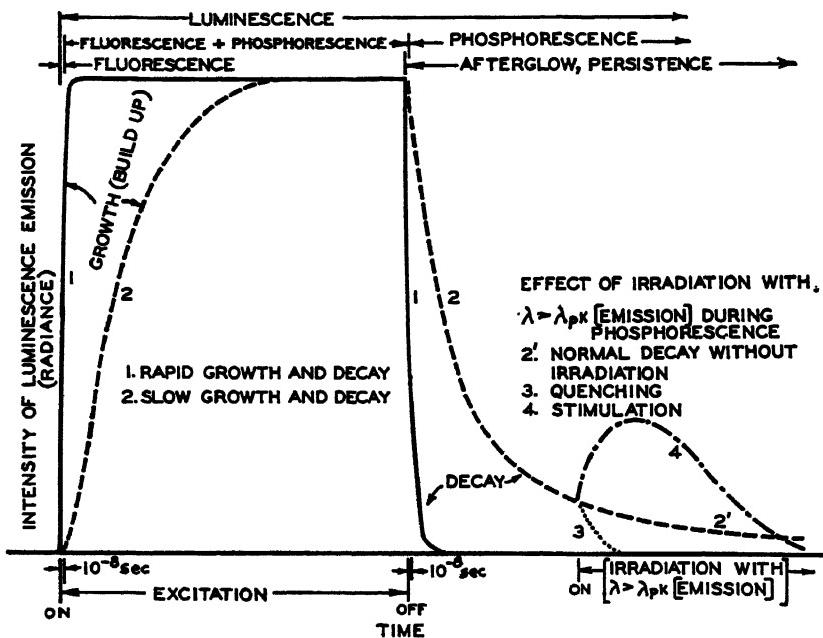
Figure 2. Diagrammatic representation of the dynamics (growth and decay) and terminology of luminescence emission.

affected by changes in the temperature of the phosphor or the conditions of excitation.

The emitting center loses control over τ when the energy storage in phosphors consists of trapped excited electrons or metastable states, for then additional *activation energy* must be supplied to release the trapped electrons. This activation energy may be supplied by heat, especially when the trap depth is equal to or less than about 30 kT, or it may be supplied by additional photons or charged material particles. During ordinary phosphorescence at room temperature the activation energy is supplied by heat, so this process is thermostimulated phosphorescence. Under these conditions, so-called *power-law decays* are observed such that $L \propto L_0 t^{-n}$, where the exponent n is

strongly dependent on the phosphor temperature and on the kind, intensity, and duration of excitation; also, n varies during the decay interval. In general, n has values lying between about 0.1 and 2, being near unity for most of the useful phosphorescence times of efficient long-persistent phosphors. The variability of n bespeaks the variable degree of filling and rate of emptying of traps of different densities and depths, and the variable occurrence of retrapping, since the density of excitation decreases in a
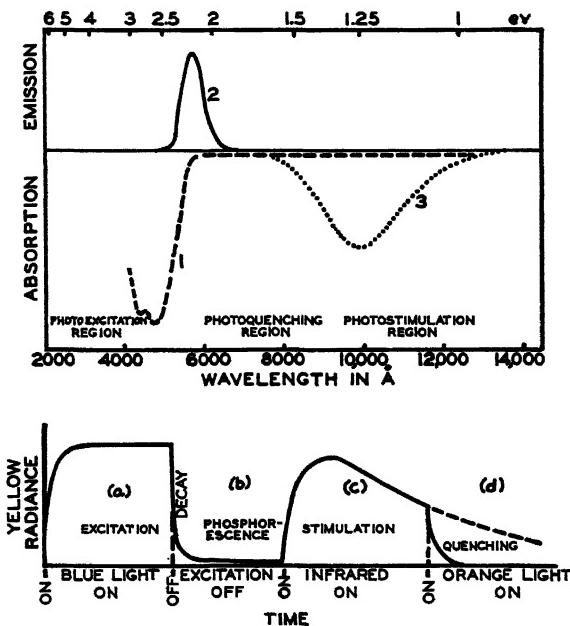
Figure 3. *Above:* Spectral distributions of excitation, quenching, and stimulation for a cub.-Sr(S:Se):SrSO₄:CaF₂:Sm:Eu phosphor. *Below:* Dynamics of growth and decay of the above phosphor during and after (a) excitation, (b) decay, (c) stimulation, and (d) quenching.

roughly exponential manner as the beam of primary excitant particles penetrates into a phosphor crystal.

Infrared is another useful source of activation energy which may (1) be unabsorbed (i.e., be entirely reflected or transmitted) and hence cause no change in the normal decay curve shown as 2′ in Figure 2, or (2) be absorbed and *quench* the phosphorescence emission, as shown in curve 3 of Figure 2, or (3) be absorbed and *stimulate* the phosphorescence emission, as shown in curve 4 of Figure 2. Because the quenching and stimulating effects for a given phosphor vary with wavelength, a broad-band source of red and infrared may act both as a quenching and stimulating agent simultaneously.

TABLE 1. COMPOSITIONS, PREPARATIONS, AND DESIGNATIONS OF SOME TYPICAL PHOSPHORS

No.	Ingredients			Cryst. Temp.	Post-cryst. Treatment	Phosphor Notation	Decay Type
	Host Crystal	Flux	Added Activator				
1.	1 ZnS (i.e., 97.44 g ZnS)	2 g NaCl	—	1250°C	wash	hex-ZnS:[Zn]	l^{-n}
2.	1 ZnS	2 g NaCl	0.017 g AgNO$_3$	"	"	hex-ZnS:Ag(0.01)	"
3.	1 ZnS	2 g NaCl	0.013 g CuCl$_2$		"	hex-ZnS:Cu(0.01)	"
4.	2 ZnO + 1 SiO$_2$ (162.76 g + 60.06 g)	—	—	"	—	rbhdl.-Zn$_2$SiO$_4$:[Si]	$\epsilon^{-at} \to l^{-n}$
5.	2 ZnO + 1.02 SiO$_2$ (162.76 g + 61.26 g)	—	0.02 TiO$_2$ (1.6 g)	"	—	rbhdl.-Zn$_2$SiO$_4$:Ti(0.4)	"
6.	2 ZnO + 1.012 SiO$_2$ (162.76 g + 60.78 g)	—	0.012 MnO (1.1 g)		—	rbhdl.-Zn$_2$SiO$_4$:Mn(0.3)	"

The top portion of Figure 3 shows the spectral relationships of excitation, emission, quenching, and stimulation for a complex infrared-stimulable phosphor of the type used in metascopes for infrared detection and signalling. The bottom portion of Figure 3 shows how the yellow emission band may be produced by excitation with blue light (≈ 4500Å), allowed to decay, stimulated by infrared ($\approx 10,000$Å,) and quenched by orange light (≈ 6000Å).

The complex chemical formula (which is really an oversimplification!) for the infrared-stimulable phosphor of Figure 3 serves as an extreme example of the system of notation used in symbolizing luminescent solids. The chief ingredients of a good phosphor are the host crystal, one or more fluxes (fusible salts, which are not always used), and one or more activators (promoters of luminescence) which may be added deliberately or be induced by decomposition during crystallization at high temperatures. Examples of other useful phosphors and their notations are given in Table 1. According to the indicated simplified notation, a phosphor is symbolized by (1) the crystal system of the host crystal, followed by (2) the chemical formula of the host crystal, then by (3) any fluxes which are *incorporated* in the host crystal (this is not the case in examples 1 to 3 of Table 1), and finally by (4) the chemical identity of the activator *atom* [placed in square brackets when there is uncertainty as to its identity or presence] with the weight per cent of activator atom relative to the weight of the host crystal given in parentheses.

Exemplary Syntheses of Phosphors

The phosphor compositions given in Table 1 are sufficient to prepare efficient phosphors *if* exceptionally *pure ingredients* are used (especially when the optimum activator proportion is low), the ingredients are *thoroughly mixed*, and the mixtures, in acid-cleaned, covered, fused-silica crucibles, are *heated* in air to the indicated temperatures for ten to a hundred minutes. These examples are typical of the formation of most phosphors by reactions in the solid state, i.e., reactions between solids at temperatures below their melting points and the melting points of their products. (Growth of phosphor crystals from melts is seldom successful because the melting points of the host crystals are usually so high that the melts react vigorously with their containers and the mixtures tend to decompose and volatilize selectively.) Under these conditions of solid-state reaction, the presence of a flux, such as sodium chloride, sometimes promotes crystallization by providing a fluid phase to facilitate material transport. Figure 4a shows an electron micrograph of a pure, practically nonluminescent, precipitated-and-dried zinc sulfide, which was used as the initial ingredient in preparing phosphors 1 to 3 of Table 1. Figure 4b shows a photomicrograph of the much larger phosphor crystals produced by heating this fine zinc sulfide (with flux) to 1250°C. These hex.-ZnS:Ag(0.015) phosphor crystals are about 10^8 times larger (in volume

(a)

(b)
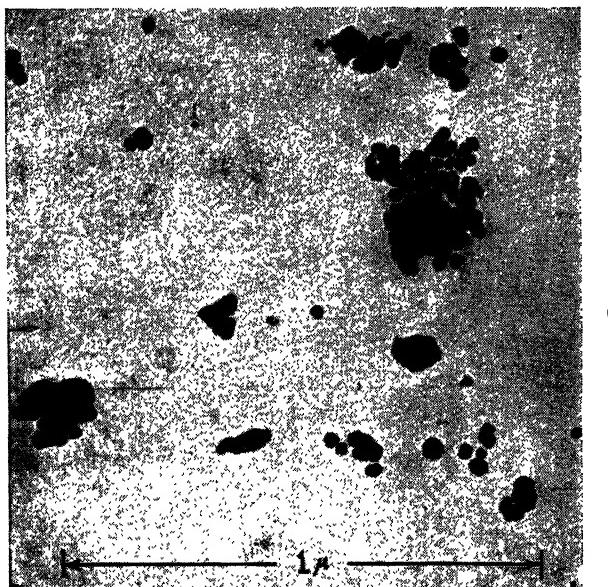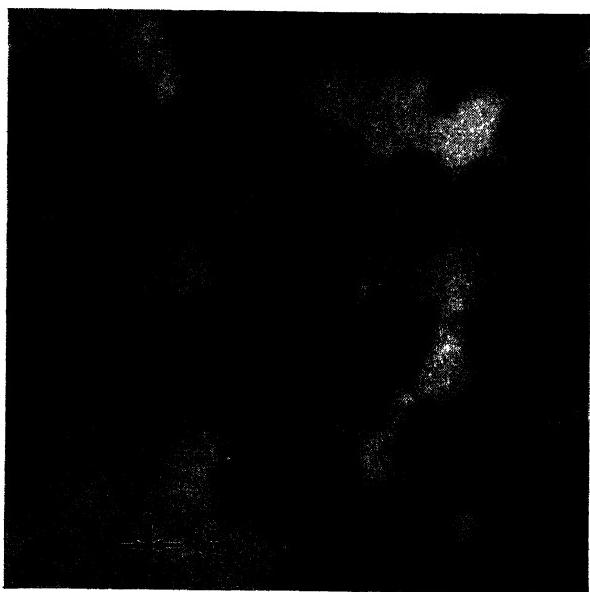
Micrographs by Dr. J. Hillier

Figure 4a. Electron micrograph of precipitated-and-dried, lumi-
nescence-pure zinc sulfide.
   Figure 4b. Photomicrograph of hex.-ZnS:Ag (0.015), prepared
by heating the fine-particle ZnS shown in Figure 4a (with acti-
vator and about 6 per cent halide flux) at 1250°C.

(a)

(b)
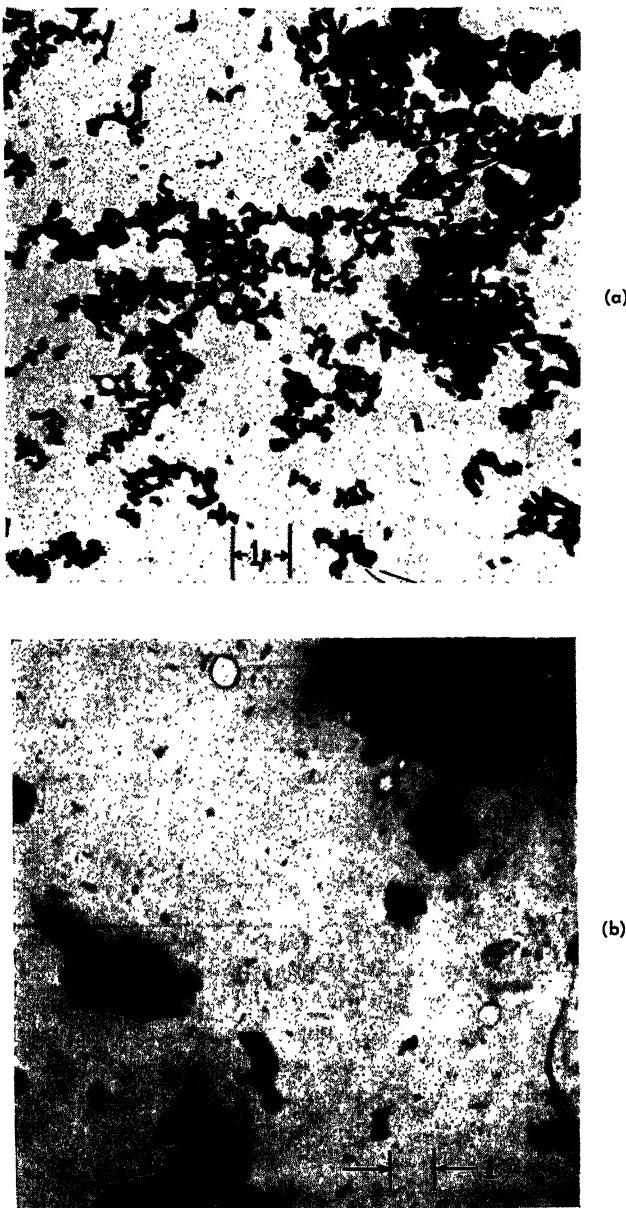
Micrographs by Dr. J. Hillier

Figure 5a. Electron micrograph of very fine particle rbhdl.-Zn$_2$SiO$_4$:Mn which was prepared by reaction of zinc hydroxide and manganese hydroxide with colloidal silica.

Figure 5b. Electron micrograph of commercial rbhdl.-Zn$_2$SiO$_4$:Mn prepared by reaction of the oxides, using conventional pure silica with an average particle size of about one micron.

and weight) than the initial 250Å ingredient particles. In the case of the zinc-silicate phosphors, no flux is used, and the average particle size of the resultant phosphor is determined largely by the particle size of the initial silica, because the reaction proceeds by diffusion of the zinc and manganese oxides into the silica particles. This is illustrated in Figure 5, which shows electron micrographs of two different lots of rbhdl.-Zn_2SiO_4:Mn(0.3) prepared from (a) exceptionally fine colloidal silica, and (b) an ordinary c.p. silica.

As may be seen from Figures 4 and 5, the average particle sizes of useful phosphors are generally a few microns or less, because finely-divided ingredients must be used to obtain complete solid-state reaction in a reasonably short time. The small particle size of phosphors is desirable to increase the absorption of primary ultraviolet in "fluorescent" lamps, and to improve the image definition and optical efficiency of television cathode-ray-tube screens, but it makes difficult the determination of certain fundamental physical characteristics such as absorption spectra, absorption coefficients, and conductivities. It is in only a few cases, such as cub.-KCl:Tl, rbhdl.-Al_2O_3:Cr (artificial ruby), and tetr.-$CaWO_4$:[W] (artificial scheelite), that efficient luminescent crystals of centimeter size have been prepared from their melts. Insofar as the effect of crystal size on luminescence is concerned, the luminescence of phosphors is a *volume* effect, as evidenced by the fact that (1) cathodoluminescence efficiency increases as the penetration of the primary electrons increases, and (2) the two different lots of rbhdl.-Zn_2SiO_4:Mn(0.3) phosphor shown in Figure 5 have practically the same efficiency of cathodoluminescence and photoluminescence despite the great difference in their particle sizes.

Effects Produced by Impurities

In general, phosphors consist of relatively nonluminescent host crystals containing a small proportion of added or induced impurities. Impurities can form many different local distorted regions (e.g., centers), as depicted in Figure 6, and a given impurity may: (1) act as an *intensifier activator* by intensifying a weak or latent host-crystal emission spectrum, (2) act as an *originative activator* by producing a new emission spectrum, (3) act as a *sensitizer* by producing a new excitation spectrum without altering the emission spectrum, (4) act as a *trap* by altering the duration and intensity of power-law-type phosphorescence, and (5) act as a *"poison"* (or "killer") by decreasing luminescence efficiency.

The following examples illustrate these effects and show how a given impurity may function in several roles:

(1) When very pure zinc silicate is crystallized at 1250°C, the product is found to have an inefficient cathodoluminescence emission shown as curve 4

in Figure 7. This emission comes from some of the excited tetrahedral SiO_4 groups which are presumably perturbed by a slight excess of silicon [Si] (or interstitial [Zn]?) produced by selective volatilization of oxygen. Here the [Si] acts as an *intensifier activator*, probably by upsetting the selection rules which govern electronic transitions in ideal crystals. When about one per cent of titania is incorporated in this phosphor, the same emission band is obtained with about a tenfold increase in efficiency (cf. curve 5, Figure 7). Here, the titanium (in the combined form!) is an additional intensifier activator. In the same way, silver appears to act as an intensifier activator for zinc-sulfide

(a)   (b)   (c)

○   HOST - CRYSTAL CATION (C)

○   HOST - CRYSTAL ANION (A)

●   CATION IMPURITY (I_C)

⊕   ANION IMPURITY (I_A)
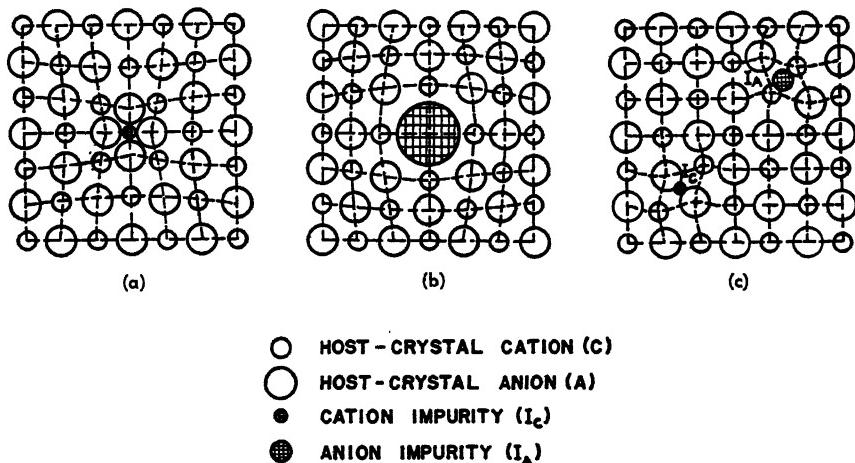
Figure 6. Examples of typical distortions (centers) produced by impurities in host crystals. The electron clouds of the atoms are generally *not* spherical, and may have distorted shapes in the centers.

(a) Small, or high-positive-charge, substitutional cation impurity (I_C). (Same effect produced by small or high-negative-charge substitutional anion.)

(b) Large, or low-negative-charge, substitutional anion impurity (I_A). (Same effect produced by large or low-positive-charge substitutional cation.)

(c) Interstitial impurities, showing effects on ligands (nearest bound neighbors).

phosphors, operating to increase the efficiency with only a slight shift and narrowing of the emission band (curves 1 and 2, Figure 7).

(2) When rbhdl.-Zn_2SiO_4:[Si] is reheated at 1250°C with increasing proportions of manganese oxide, the [Si] emission band decreases and a new emission band rises in the green until, at about one per cent Mn, only the strong green emission is evident (curve 6, Figure 7). Here, the Mn (which substitutes for Zn) acts as a *poison* in preventing the [Si] emission, and simultaneously acts as an *originative activator* in producing the new centers which emit luminescence photons with high efficiency. Similarly, copper activator poisons the blue [Zn] emission band of hex.-ZnS:[Zn] while producing a new green emission band (curve 3, Figure 7). Present evidence

indicates that the best originative activators are added ions which exhibit more than one formal valence (e.g., Mn^{++}, Mn^{++++} and Cu^{+}, Cu^{++}), whereas intensifier activators may be either added or induced, and need not be multivalent (e.g., $[Zn^{++}]$, $[Si^{++++}]$; Ti^{+++}, Ti^{++++}; and Ag^{+}, Ag^{++}).

(3) When calcium carbonate or calcium silicate is crystallized with about one per cent of manganese activator, the phosphors are found to have efficient red-orange cathodoluminescence, but no appreciable photo-luminescence under 2537Å ultraviolet. If, however, these phosphors are recrystallized with a small proportion of lead oxide in addition to the added manganese oxide, then the same red-orange emission is obtained as before and the phosphors are now readily excited by 2537Å ultraviolet. Here the
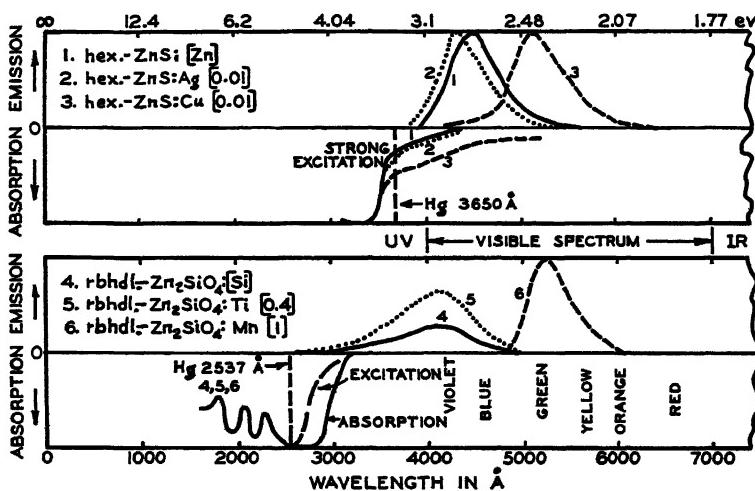
Figure 7. Spectral distribution of absorption, excitation, and luminescence emission of some typical zinc-sulfide-type and zinc-silicate-type phosphors.

(combined) lead acts as a *sensitizer* which introduces a new absorption band in the spectral region around 2537Å.

(4) In the previous example of hex.-ZnS:Cu, the multivalent Cu acts not only as a poison and an activator but also as a *trapping agent*. This is evidenced by intense and prolonged phosphorescence at room temperature and by pronounced glow curves (curves of light output vs. temperature and time as a phosphor previously excited at very low temperatures is warmed). By incorporating a lead compound in this phosphor, it is found that the lead provides very deep traps, such that the phosphor performs in the same manner as the infrared-stimulable material in Figure 3. In the case of rbhdl.-Zn_2SiO_4:Mn, it has been found that tin and arsenic compounds are effective in providing traps without altering the emission spectrum of the phosphor.

(5) In addition to the previous examples of poisons, it may be mentioned that manganese in trace amounts is a strong *poison* to the emission of calcium-tungstate phosphors. Also, certain other transition elements, particularly iron and cobalt, are strong poisons in almost all phosphors.

Effects Produced by Changes in Structure

The pronounced influence of structural changes on the luminescence of solids is shown in the following examples:

(1) When rbhdl.-Zn_2SiO_4:Mn (denoted as the α-form) is melted at 1600°C and then quenched, it crystallizes in a new (β-form; undetermined) structure having a different diffraction pattern. On comparing the α and β products, it is found that: (1) the emission band of the α-form peaks at 5250Å, whereas the emission band of the β-form peaks at 5630Å; (2) the β-form has about 75 per cent of the cathodoluminescence and photo-luminescence efficiency of the α-form; and (3) the α- and β-forms have practically identical exponential decay characteristics. On heating the yellow-emitting β-form to about 950°C, it reverts to the green-emitting α-form, demonstrating that the observed effects are caused by changes in structure rather than composition.

(2) When ZnS:[NaCl(2)]:Ag(0.01) is heated at 780°C it crystallizes in the cubic system, whereas on heating at 1200°C it crystallizes in the hexagonal system. On comparing these two products, it is found that: (1) the emission band of the cubic material peaks at 4480Å, whereas the emission band of the hexagonal material peaks at 4330Å; (2) the hexagonal material has about four times the peak output of both cathodo- and photo-luminescence relative to the cubic material; and (3) although both materials have power-law decays, the intensity and duration of phosphorescence emission of the hexagonal material is much greater than that of the cubic material at room temperature. On grinding, the hexagonal crystals are transformed into the cubic structure.

Luminescence Excitation

When a beam of primary excitant particles impinges upon a phosphor crystal, some particles are reflected, whereas others are transmitted into the interior of the crystal. The energy of a transmitted primary *photon* is absorbed all at once or not at all (neglecting Compton scattering) in one absorption act wherein the photon is annihilated. The energy of a fast primary *electron* or *ion*, on the other hand, is usually absorbed bit-wise, where the average energy bit is about 25 ev. Differences in excitation processes are evident, also, with respect to *where* the energy is absorbed in a phosphor crystal. Low-energy primary ultraviolet photons often excite phosphor centers directly, because the centers generally have lower characteristic frequencies than the atoms of

the host crystal. High-energy primary particles (e.g., x-ray and gamma-ray photons, and fast electrons and ions), however, give up their energies indiscriminately, so that most of their energy is absorbed by the preponderant host-crystal atoms. In this case, the absorbed energy must be transported to the centers and the efficiency of such energy transport depends strongly on the degree of perfection of the crystal. Hence, there are many photoluminescent glasses and crystals with high efficiency, but only highly crystalline phosphors give high efficiency of roentgenoluminescence, cathodoluminescence, and ionoluminescence. The high rate of degradation (into heat) of energy transported in vitreous matter accounts, also, for the fact that only
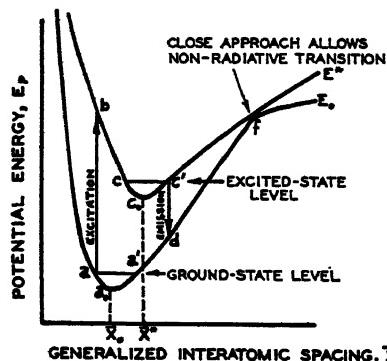
Figure 8. Configuration-coordinate energy-level diagram of a phosphor impurity center, where the potential energy of the center is plotted as a function of the average distance between the atoms in the center.

crystalline phosphors give efficient, long-persistent, power-law-decay phosphorescence involving wandering and remote trapping of excited electrons.

It is probable that most of the energy transport in phosphors is by excited "free" electrons, although other means are: positive holes, excitons, photons, and exchange-type energy transfers which occur when the wave functions of atoms and centers overlap in a suitable manner. Positive holes are residual regions of excess positive charge produced by internal ionization, and excitons are mobile pairs of positive holes and nearly-"free" excited electrons. The problems of energy transport in solids are complicated by the interplay of ionic (electrostatic) and covalent (shared-electron) bonding between the atoms, and by the pronounced influences of imperfections and different structural arrangements in crystals.

Luminescence Mechanism

For lack of specific information about the energy levels in phosphors, two complementary types of simplified energy-level diagrams have been devised to give a generalized picture of the mechanism of luminescence. One such diagram, shown in Figure 8, depicts the allowed potential energies of a

luminescence center as a function of averaged interatomic spacing, $\bar{x}$, where $\bar{x}_0$ is the averaged spacing between the atoms of the unexcited center at equilibrium. At 0°K, the ground-state energy level would be very near a_0, corresponding to minimum potential energy and minimum atomic vibration. At room temperature, however, the system (center) has considerable vibrational energy, so the ground-state level lies higher, such as at a, where the amplitude of atomic vibration is proportional to $a' - a$ (imagine the system passing through a_0 as it rolls from a to a' and back, in the potential well). When a bit of excitation energy of adequate magnitude, usually greater than 2 ev, is transmitted to the center, the energy of the center may be raised from a on the ground-state curve, E_0, to b on the excited-state curve, E^*. Within about 10^{-12} second after excitation, the atoms of the excited center readjust to a new equilibrium spacing, $\bar{x}^*$, and the energy difference $b - c$ is given up as heat to the surrounding host crystal. With the center in the excited-state level c, the probability of a radiative transition from c' to d is determined by the natures of the impurity and host crystal, being practically independent of temperature, but the probability of a nonradiative transition to the ground state via $c' \rightarrow f \rightarrow a$ increases exponentially with temperature. Here, then, the observed phosphorescence for this luminescence without internal ionization is practically independent of the temperature when $kT << \Delta E$, or of the kind, duration, and intensity of excitation; and the decay proceeds according to

$$L = L_0 \epsilon^{-(\alpha + \nu_a \epsilon^{-\Delta E/kT})t}$$

where the atomic vibration frequency, ν_a, is about 10^{12} sec^{-1}, and the thermal activation energy, ΔE, is the energy difference $f - c_0$. The observed decrease of luminescence efficiency with increasing temperature may be visualized as a raising of level c so that an increasing proportion of excited states "spill over" via f without producing radiation. If, at any given temperatures, the levels c and f practically coincide, then the center is a *poison* in that it is a means for rapidly degrading excitation energy into heat. This picture, then, illustrates how a center which produces luminescence at low temperatures may be a poison center at higher temperatures. The number of available efficient phosphors decreases rapidly with increasing operating temperature, with hardly any phosphors having useful efficiencies above 400°C. In general, phosphors should be operated at as low temperatures as possible, although some power-law-decay phosphors exhibit an intermediate optimum operating temperature where traps are efficiently emptied by thermal energy.

Another type of energy-level diagram, shown in Figure 9, depicts the energy levels of a luminescence center as a function of distance along a row of atoms in the crystal. This type of diagram emphasizes the fact that the discrete energy levels of isolated atoms are spread out into bands in solids, the broadening being caused by interaction of the electrons of all the atoms

in the crystal, because the Pauli exclusion principle allows only two electrons (of opposite spin) to occupy the same energy level in a given system. The diagram is drawn for a specific $\bar{x}$, so one must imagine that the spacings and potential barriers between the atoms change with every energy change
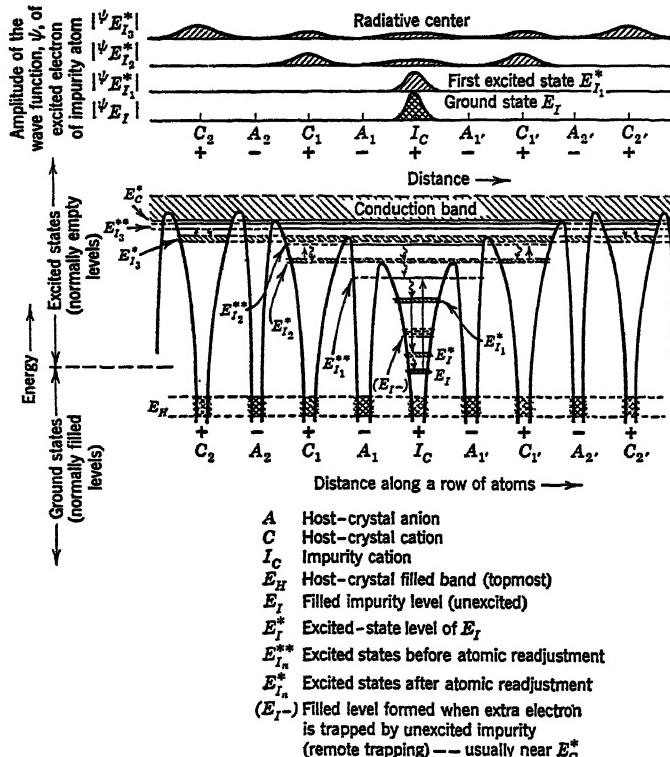
Figure 9. Energy-level diagram of a row of atoms in a phosphor impurity center. The lower part of the figure shows the lowering of the potential barriers in the neighborhood of the impurity, and the relatively discrete occupied and excited-state energy levels introduced by the impurity. The upper part of the figure shows a plot of the absolute value of the wave function, ψ, of the optical electron responsible for luminescence (ψ^2 is a measure of the probability of finding the electron in a given location). The breadths of the excited-state levels increase with height and so the upper levels may overlap each other and the conduction band.

of the center. According to the diagram, the impurity atom, I_C: (1) lowers the normal potential barriers between the host-crystal atoms in its vicinity, (2) introduces an additional occupied level, E_I, into the forbidden zone of host-crystal energies, and (3) introduces additional discrete unoccupied excited-state levels, $E_{I_n}^*$, whose extensions into the surrounding crystal

increase as $E_{I_n}^*$ increases until, in the conduction band, E_C^*, an excited electron is free to move through the host crystal. The excitation transition $E_I \to E_{I_i}^{**}$ in Figure 9 corresponds to $a \to b$ in Figure 8, and the radiative return $E_{I_1}^* \to E_I^*$ corresponds to $c' \to d$. As drawn, this process is highly localized, and the spontaneous decay is exponential. When the low-lying, highly localized excited state, $E_{I_1}^*$, is absent, and the center is excited to higher levels wherein the excited electron may wander some distance away from its parent atom, then the electron may become trapped at a distant point, where it must be released by an additional activation energy before it can return to make a radiative transition from a low-lying $E_{I_n}^*$ level. When the excited electron travels through the conduction band, E_C^*, it may be trapped by a remote unexcited center, forming a new filled level (e.g., E_{I}-). This type of trapping is most probable when the impurity is a multivalent ion; for example, a Sm^{+++} impurity can trap an extra electron to become Sm^{++}. Regardless of whether the excited electron is trapped in a metastable state in its center of origin or in a remote center, the phosphorescence is strongly dependent on temperature and on the kind, duration, and intensity of excitation. The complex decay then proceeds by thermally activated release of trapped electrons according to

$$L = L_{0_1}\epsilon^{-\nu_{a_1}t_\epsilon - \Delta E_1/kT} + L_{0_2}\epsilon^{-\nu_{a_2}t_\epsilon - \Delta E_2/kT} \cdots + L_{0_N}\epsilon^{-\nu_{a_N}t_\epsilon - \Delta E_N/kT}$$

where the subscripts $1, 2, 3 \cdots N$ denote traps of different densities and depths ΔE_1, ΔE_2, ΔE_3, $\cdots \Delta E_N$ which make different contributions L_{0_1}, L_{0_2}, $L_{0_3} \cdots$, L_{0_N} to the luminescence output at time $t = 0$. This lengthy expression is usually shorthanded by the power-law-decay *approximation* $L = L_0 t^{-n}$, since, as previously noted, n varies not only with the type of phosphor, but with the temperature, decay time, and conditions of operation. Both exponential and power-law decays can occur, without change in spectral distribution, by exciting a center such as the one generalized in Figure 9 to $E_{I_1}^{**}$ and higher levels, as long as the final radiative transitions take place from $E_{I_1}^*$ to E_I^*. Almost all phosphors which exhibit initial exponential decays eventually trail off into power-law "tails" which bespeak the delayed returns of distant trapped electrons that make the same final radiative transition as do the highly localized untrapped excited electrons responsible for the predominant exponential decay. As is to be expected from Figure 9, electronic conduction "parallels" the growth and decay of many power-law-decay phosphors, whereas little or no conduction is observed when the decay is predominantly exponential.

Stimulation of phosphorescence by low-energy (long-wavelength) photons, as in Figure 3, is readily understood by picturing the weak photons as having just enough energy to raise trapped electrons out of their traps so they can make radiative recombinations with the parent ionized centers. Quenching of phosphorescence by higher-energy photons, however, apparently involves

exciting the system to a high-energy level near or above f in Figure 8, i.e., the trapped excited electron is raised so high in energy that the surrounding atoms are set in violent agitation and the excitation energy is dissipated as heat.

Luminescence Emission Spectra

Both line and band emission spectra may be produced separately or simultaneously by phosphors. Line spectra are obtained when radiative transitions take place between discrete highly localized energy levels. Such discrete levels often occur when well-oxidized impurity ions with incomplete inner shells (e.g., Eu^{+++} and Cr^{+++}) form luminescence centers. For example, $ThSiO_4$:Eu(1), crystallized at 1250°C in oxygen, gives a line emission spectrum attributable to electrons with unpaired spins in the *inner*, well-shielded, incomplete $4f$ shell of Eu^{+++} making transitions between discrete subatomic levels having different resultant spin quantum numbers. When this phosphor is heated in a reducing atmosphere, however, the unpaired $4f$-shell spins are apparently paired and a band emission is obtained in place of the previous line emission. The band emission is attributed to the *outer* valence electrons of Eu^{++} making transitions between levels having different principal and angular-momentum quantum numbers (where at least one of the levels is a band).

The chemical and structural constitutions of phosphors determine, of course, the locations and breadths of their emission lines and bands. Of about 10^5 different samples of artificial inorganic phosphors that have been synthesized here and abroad, the most useful phosphors have band widths of the order of 0.7 ev, although lines narrower than 10^{-3} ev and bands broader than 2 ev have been obtained in some cases. Empirical methods for controlling the locations of emission bands in almost any part of the visible and near-visible spectrum have been developed for several efficient phosphor families where variations in host-crystal composition, in particular, afford excellent control over the properties of the resultant phosphors. For example, increasing partial substitution of cadmium for zinc (or selenium for sulfur) in any of the three zinc-sulfide phosphors shown in Figure 7 gradually shifts the emission spectrum toward the infrared. At present, the white-emitting luminescent screens of direct-viewing television cathode-ray tubes are made of a mixture of blue-emitting hex.-ZnS:Ag(0.01) and complementary yellow-emitting hex.-1.3ZnS·CdS:Ag(0.01). (The latter phosphor is also useful for x-ray fluoroscope screens.) Similarly, luminescent screens for direct-viewing color television may be obtained by selecting appropriate blue-, green-, and red-emitting members of this phosphor family. In the case of the outstandingly useful phosphor family based on rbhdl.-Zn_2SiO_4:Mn, increasing substitution of germanium for silicon

produces a gradual shift of the emission band toward the red, but when beryllium is substituted in increasing proportions for the zinc, the original green emission band peaked at 5250Å decreases and a new orange-red emission band rises at 6300Å. The high activator proportions of silicate phosphors make them particularly useful for operation under conditions of intense excitation. For example, the three different luminescent screens for the three (trinoscope) cathode-ray tubes used in laboratory demonstrations of color television on a theatre scale may be made of blue-emitting rbhdl.-Zn_2SiO_4:Ti(1), green-emitting rbhdl.-Zn_2SiO_4:Mn(1), and red-emitting rbhdl.-$Zn_8BeSi_5O_{19}$:Mn(2). White-emitting luminescent screens for "fluorescent" lamps are made either by mixing blue-emitting monocl.-Mg_2WO_5:[W] with yellow-emitting rbhdl.-$(Zn:Be)_2SiO_4$:Mn, or by using a calcium-fluoro(chloro)-phosphate:Mn:Sb phosphor whose two different activators produce two complementary emission bands. It may be noted that the influence of new host-crystal atoms should depend not only on the nature of the new atom, but also on whether the new atom becomes an immediate neighbor or just a near neighbor of the emitting atom or center, and whether the luminescent center is founded on a substitutional or interstitial impurity.

Luminescence Efficiency

When phosphors are excited by low-energy photons (*photoluminescence*), quantum efficiencies exceeding 90 per cent have sometimes been obtained. For example, the phosphor coating of a green-emitting "fluorescent" lamp converts over 90 per cent of the input 2537Å (4.9-ev) primary photons into emitted luminescence photons having an average energy of about 2.4 ev (5250Å). On an energy basis, this luminescence process is $(2.4/4.9)90 \approx 45$ per cent efficient. In some cases, the energies of the primary and emitted photons lie even closer together, so that the energy efficiency may be higher. It is rare, however, that the energy efficiency of photoluminescence of phosphors exceeds about 80 per cent, because there is always some gap between the peaks of their excitation and emission bands (cf. Figure 7). For a given average energy of the emitted photons $h\bar{\nu}_{em}$, the energy efficiency, $\mathcal{E}$, decreases with increasing energy of the excitant photons, $h\bar{\nu}_{ex}$; i.e., $\mathcal{E} \propto \bar{\nu}_{em}/\bar{\nu}_{ex}$. This is true until $\bar{\nu}_{ex}$ becomes large enough to produce internal photoelectrons with sufficient energy to produce in turn more than one luminescence photon. When this happens, as it does for excitation by x-rays and gamma rays (*roentgenoluminescence*), the luminescence process is essentially cathodoluminescence.

When phosphors are excited by fast charged material particles (*cathodoluminescence* and *ionoluminescence*), the energy efficiency is vanishingly small for low primary-particle energies because slow particles dissipate their energies

in the inefficient (distorted and chemically different) surface layers of the phosphor crystals. It is only in rare cases that detectable luminescence output can be obtained from phosphors excited by 5-ev primary electrons, and even at primary energies of thousands of volts the energy efficiency does not exceed about 10 per cent. The low efficiency which obtains even when the primary particles penetrate well into the efficient volumes of phosphor crystals is attributed to the large difference between the absorbed energy bits (approximately 25-ev average) and the emitted photons (approximately 2.5-ev) and to the difficulties of energy transfer from the predominant host-crystal absorber atoms to the fewer activator centers.

Luminescence Output

The radiances of phosphors *during* excitation by photons are limited mostly by the low intensities of available photon sources. With available sources of ultraviolet, for example, the maximum luminances of visible-light-emitting phosphor screens are of the order of 5000 millilamberts (mL), or about 2×10^{16} photons emitted per square centimeter per second. On the other hand, brief instantaneous luminances exceeding 10^7 mL (over 10^{20} photons $cm^{-2} sec^{-1}$) can be obtained from phosphor screens struck by a well-focused, high-voltage, scanning cathode-ray beam, although the sustained averaged luminance of such a scanned screen generally may not exceed about 10^5 mL (4×10^{17} photons $cm^{-2} sec^{-1}$) without heating the screen above the temperature range in which appreciable luminescence efficiency is obtained. (It may be recalled that fresh snow in full sunlight has a painful luminance of about 10^4 mL.) In a conventional direct-viewing television cathode-ray tube, producing 50 to 100 mL in the image highlights, a given screen element the size of the cathode-ray-beam area is excited for about 1.5×10^{-7} second, 30 times a second, for a summed duration of only 16 seconds of actual operation time per 1000 hours of total operation time. On this basis alone, it seems odd that the efficiency of the screen should change somewhat during 1000 hours elapsed (16 seconds actual) time of operation. During the actual operation time, however, the instantaneous power input into a screen element is about 10,000 volts $\times 2 \times 10^{-4}$ ampere/0.001 cm^2 = 2000 watts/cm^2. This power loading is almost ten times as high as the 250 watts/cm^2 absorbed power input (and radiated power output) of a tungsten filament in an ordinary incandescent lamp. In projection cathode-ray tubes, the power loadings may be several orders of magnitude higher than in direct-viewing cathode-ray tubes.

The radiances of phosphors *after* excitation are determined by their decay characteristics and light sums under the operating conditions. The *light sum* is the total radiance per unit area integrated over the entire afterglow interval. Integration of the decay curves of trap-type phosphors such as hex.-ZnS:Cu and cub.-Sr(S:Se):SrSO$_4$:CaF$_2$:Sm:Eu (taking into

account the penetration of the primary particles) has shown that about 10^{18} potential photons/cm^3 may be stored in the excited volumes of some phosphor crystals under favorable conditions. This number is of the same order of magnitude as the number of Cu or Sm impurity centers in the cited phosphors. To a first approximation, the light sum is proportional to (1) the number of excitable centers and traps per unit volume, (2) the depth of penetration of the primary particles, and (3) the intensity and duration of the excitation (up to saturation).

The time interval in which the light sum is practically all released may be varied enormously by careful choice of phosphors and operating conditions. For example, (1) screens of hex.-ZnO:[Zn], cub.-MgS:Sb, and rhomb.-BaSO$_4$:Pb give up most of their stored luminescence energy in about 10^{-6} second in cathode-ray tubes used for flying-spot image pickup in television, (2) a cascade screen, wherein a cathodoluminescent blue-emitting hex.-ZnS:Ag phosphor excites a photoluminescent yellow-emitting hex.-9ZnS·CdS:Cu phosphor, emits about 0.03 mL three seconds after excitation in radar cathode-ray tubes, and has luminances which are detectable by the well-dark-adapted eye for many hours, and (3) screens of photoluminescent cub.-Sr(S:Se):SrSO$_4$:CaF$_2$:Sm:Eu retain most of their light sums for about six months at room temperature, thereby providing useful long-duration retentivity of information and infrared stimulability.

Concluding Remarks

Man-made phosphors, which were unimportant alchemical novelties during the 17th and 18th centuries, found their first important uses in the 19th century as visible indicators of certain invisible energetic particles, such as ultraviolet and x-ray photons, cathode rays, and alpha particles. In recent years of the 20th century, the *direct* conversion of the energies of these invisible particles into light, at operating temperatures near room temperature, has become a major commercial function of phosphors which are now produced at the rate of over 200,000 kilograms a year. Meanwhile, phosphors are finding increasing scientific use in detecting these invisible particles and others, including infrared- and gamma-ray photons, fast-moving ions, and even neutrons, by converting their energies into radiations which the human eye may detect directly, or indirectly through other photosensitive devices, such as multiplier phototubes (usually coupled with oscilloscopes or meters), photographic films, or other phosphors used in cascade.

In addition to the practical progress already made by empirical phosphor research, some progress has been made toward developing a qualitative theory of luminescence of solids, although a useful quantitative theory is not yet available. Luminescence is such a convenient and sensitive indicator of changes of composition, structure, and atomic interactions in solids that it

has contributed much to our improved understanding of the solid state of matter. In the future, the practical consequences of this broad aspect of luminescence research may well overshadow the tangible results already obtained.

REFERENCES

1. Leverenz, H. W., "Luminescent Solids (Phosphors)," *Science*, 109 (2826), 183-95 (Feb. 25, 1949).
2. Strobl, A., "Luminescent Paints, Pigments, and Inks," in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. VI, New York, Reinhold Publishing Corp., 1946.
3. Fonda, G. R., and Seitz, F., eds., "Preparation and Characteristics of Solid Luminescent Materials," New York, John Wiley & Sons, Inc., 1948.
4. Forsythe, W. E., and Adams, E. Q., "Fluorescent and Other Gaseous Discharge Lamps," New York, Murray Hill Books, Inc., 1948.
5. Kröger, F. A., "Some Aspects of the Luminescence of Solids," New York, Elsevier Publishing Co., Inc., 1948.
6. Leverenz, H. W., "An Introduction to Luminescence of Solids," New York, John Wiley & Sons, Inc., 1950.

NEWTONIAN FLOW IN POLYMER CHEMISTRY

Arthur K. Doolittle

Asst. Director of Research, Carbide and Carbon Chemicals Division,
Union Carbide and Carbon Corp., S. Charleston, W. Va.

Introduction

The scientific study of the viscosity of liquids has been in progress for over one hundred years, but the importance of the subject has been greatly emphasized in recent years because of its usefulness in connection with both the manufacture and the application of resinous substances. To mention but a few such instances, viscosity is used in the control of quality in paints, varnishes, lacquers, inks, and adhesives, for maintenance of uniformity in the manufacture of films, fibers, and coated goods, and in the control of the manufacture of the resins themselves. The viscosity effects encountered are sometimes simple, but more often they are complex and therefore difficult to define in general terms. For this reason, a great body of empirical relationships has been accumulated over the years, each of which may be useful for the particular purpose for which it was developed, but none of which applies over an extended range of the variables. In this chapter are discussed the results of recent work, which make possible considerable extension of the useful range of the variables beyond those attainable with previous relationships.

The simplest viscosity effect, known as Newtonian flow, is that in which the rate of deformation is directly proportional to the magnitude of the applied force; all other viscosity effects are superimposed on the Newtonian or fundamental condition of flow. Pure liquids and dilute polymer solutions generally exhibit Newtonian flow. Concentrated solutions of many polymers and most plastic compositions, as a rule, exhibit elastic as well as viscous effects which considerably complicate their study. The discussion in this chapter will be limited to the consideration of Newtonian flow only.

Lest the reader be misled into believing that all complications

149

will be avoided by restricting our discussion to Newtonian flow, we hasten to point out that even this simplest classification of the subject is complicated in the case of solutions by uncertainties in regard to molecular weight and molecular-weight distribution, and in the case of pure liquids, by association, particularly in the neighborhood of the freezing points. It is, perhaps, these very complications that have so greatly confused the subject; in spite of over one hundred years of active study by scientists all over the world, no completely satisfactory theoretical concept of the flow process has yet been proposed.

Empirical Viscosity Functions

Many empirical expressions defining viscosity in Newtonian flow as a function of temperature, molecular weight (in homologous series) or concentration (in solutions) have been proposed, and it will be informative to consider a number of these relations chronologically.

Formulas from the Literature for the Dependence of Viscosity on Temperature, Free-Space, Molecular Weight (in Homologous Series) and Concentration (in Solutions):

Viscosity as a Function of Temperature. The viscosity-temperature functions proposed by various investigators through the years fall into the following four general classifications: (1) binomial expansions, (2) reduced temperature expressions, (3) fluidity relations, and (4) exponential and double exponential relationships.

The formulas of Poiseuille[1] (1846), Slotte[2] (1892) and many others fall under the first classification since they can all be reduced to the following form:

$$\eta = \eta_0(1 + at)^n$$

If η be taken as -2, the above form becomes the original Poiseuille equation:

$$\eta = \frac{\eta_0}{1 + \alpha t + \beta t^2}$$

Terms higher than the square term do not seem to improve the fit of the binomial expansion formulas.

Formulas falling in the second classification, reduced temperature expressions, are limited in their range of applicability because neither critical temperatures nor boiling temperatures can be obtained experimentally with compounds of moderate to high molecular weight. In reviewing the reduced temperature relationships proposed in the literature, the work of Porter[3] (1912), Nissan[4] (1940 *et seq.*), and Telang[5] (1946) should be mentioned.

Porter uses a standard liquid such as water, and compares the viscosity of the liquid under study to that of the standard liquid. If a series of temperatures, T, are chosen at which the liquid has the same viscosity as the standard liquid at some other temperatures, T_s, then T/T_s versus T will define a straight line. From this, Porter deduces that $f(\eta) = A + B/T$, where A and B are constants characteristic of each liquid. While this particular reduced temperature scheme does not involve the experimental determination of either critical or boiling temperatures, it nevertheless is of limited value because the nature of $f(\eta)$ is not characterized.

Nissan and his associates have contributed extensively to the literature on viscosity. Their reduced temperature is the quotient T/T_b, in which T_b is the absolute temperature of the boiling point. Over moderate ranges of temperature, the Nissan relationship gives quite good agreement, but it does not appear to be any better than the Andrade function which is discussed later.

Telang's equation makes use of the difference in temperature between the critical temperature, T_c, and the absolute temperature of the measurement, T. The agreement is good for ten unassociated liquids over a limited range of temperatures. The equation is:

$$1/\eta = \frac{A}{(T_c - T)^{0.3}} - B$$

Fluidities rather than viscosities attained a certain vogue several years ago as a consequence of the supposition that the fluidities of the components of mixtures were additive. The fluidity concept was first introduced by Bingham [6] in 1910, and he and his students published extensively on this subject. Bingham's first formula was:

$$t = A\phi - B/\phi + C$$

Here, t is the temperature of measurement, ϕ is the reciprocal viscosity or fluidity, and A, B, and C are constants. A more recent fluidity equation (Bingham and Stookey,[7] 1939) is:

$$\phi/T = A + BT$$

There appears to be no advantage to be derived from the employment of fluidities rather than viscosities when a considerable range of temperatures is involved.

By far the greatest activity on the part of investigators has been in the fourth category mentioned above, that of exponential relationships. The exponential type of function seems particularly well-suited to handling viscosity data and affords practically limitless opportunities for defining fairly close fits.

The simplest exponential formula, and probably the most accurate as well, was first proposed by de Guzman [8] in 1913. This formula has been

independently rediscovered by numerous investigators, many of whom were quite unaware of prior publications proposing the same relationship. Professor Andrade, in England, seems to have done more work with the exponential reciprocal temperature relationship than any other investigator, and he was the first to offer a quasi-theoretical explanation of the flow process based on this function. Because of the extensive contributions of Andrade to this subject, the relation first discovered by de Guzman is now generally called the Andrade function and this practice will be continued by the present author.

The principal simple exponential viscosity-temperature relationships found in the literature are listed chronologically below. Only those marked by an asterisk differ materially from the Andrade function.

Date	Investigator	Relationship
1913	de Guzman [8]	$\dfrac{d \ln \phi}{dT} = \dfrac{W}{RT^2}$
1916	Arrhenius [9]	$\dfrac{d \ln \eta \cdot v^{1/2}}{dT} = \dfrac{K}{T^2}$
1917	Kendall and Monroe [10]	$\ln \eta^{1/3} = \dfrac{B}{T} + a$
1918	Drucker [11]	Drucker and de Guzman collaborated on $\dfrac{d \ln \phi}{dT} = \dfrac{W}{RT^2}$
1921	Vogel [12]	$\eta = \eta_\infty{}^{(t-a)/(t-b)}$
1923	Raman [13]	$\eta_l = \eta_v \, e^{(E_2 - E_1)/RT}$
1926	Frenkel [14]	$\eta = ATe^{B/T}$ *
1926	Dunn [15]	$\eta = Ae^{Q/RT}$
1930	Andrade [16]	$\eta = Ae^{B/T}$
1930	Madge [17]	$\eta = \dfrac{Ae^{BT}}{T - b}$ *
1930	Sheppard [18]	$\eta = Ae^{B/T}$
1933	Silverman [19]	$\eta = \dfrac{KV}{\sqrt{T}} e^{(Q/RT) - CT}$ *
1933	Cragoe [20]	$A + \ln \eta = \dfrac{C}{B + t}$
1936	Macleod [21]	$\eta = \dfrac{Be^{C/T_0}}{T}$ *
1936	Eyring [22]	$\eta = \dfrac{Nh}{V} \cdot e^{E_0/kT}$
1937	Barr [23]	$(\nu + a) = e^{(A + B/T)^C}$ *
1939	Hugel [24]	$\eta = Ae^{Q/R(T - B)}$ *
1941	Eyring [25]	$\eta = 1.09 \cdot 10^{-3} \dfrac{M^{1/2} \, T^{3/2}}{V^{2/3} \, E_{vap}} \cdot e^{E/RT}$
1947	Stuart [26]	$\eta = A - B \log (T - T_f)$ *

In the above formulas the following symbols are employed:

ϕ = fluidity = reciprocal viscosity
T = absolute temperature
η = viscosity; η_l = viscosity of liquid; η_v = viscosity of vapor
t = temperature
E, W, Q = energy or heat
T_f = absolute temperature of fusion
ν = kinematic viscosity
v = specific volume; V = molar volume
$A, a, B, b, C, k, K, \eta_\infty, R, N, h$ = constants

A number of these investigators have attempted to justify their formulas on theoretical or quasi-theoretical grounds. Since, however, no simple exponential function will reproduce the measured viscosities of unassociated liquids over an extended range, it is doubtful whether a sufficient experimental basis for these theoretical treatments has been established.

A double exponential type of function extends considerably the range over which it is possible to obtain a good fit of the data. This applies particularly well to lubricating oils at the higher temperatures of measurement. As a consequence of their success with lubricating oils, the double exponential equations have been used quite widely in industry. The American Society for Testing Materials, for example, issued a special cross-section paper plotted according to one of these equations (see below); this paper is very useful for rectifying viscosity-temperature data over considerable ranges.

The double exponential type of equations was pioneered by Walther in Germany. His first formula, proposed in 1928,[27] was:

$$\log \log \nu = -\frac{n(t - 50)}{100} + \log \log \nu_{50}$$

In 1930,[28] he proposed:

$$\log \log \nu = -n(\log T - 2.309) + \log \log \nu_{50}$$

and in 1931:[29]

$$\log \log (\nu_T + 0.95) = -n \log T/T_1 + \log \log (\nu_{T_1} + 0.95)$$

The final Walther equation, which was adopted by the ASTM for their special cross-section paper for use in plotting viscosity-temperature data, was published in 1933 [30] as:

$$\log \log (\nu + 0.8) = -n \log T/T_1 + \log \log (\nu_{T_1} + 0.8)$$

In these formulas:

ν = kinematic viscosity

t = temperature, °C

T = absolute temperature

a, n = constants

The double exponential may be better visualized by transforming it to the following form:

$$(\nu + a)^{T^n} = K$$

from which it appears that the fall in viscosity owing to a rise in temperature is compensated for by raising the value of the viscosity (plus a constant) to the T^n power.

Viscosity as a Function of Free-Space. The principal investigators who have worked with the concept of viscosity as a function of free-space are:

Date	Investigator	Relationship
1913	Batschinski [31]	$\eta = \dfrac{c}{v - b}$
1918	van der Waals, Jr. [32]	$\eta = 0.335 n^2 d^4 \sqrt{2kT/m}\, \dfrac{v}{v - b}\, e^{-E/RT}$
1923	Macleod [33]	$\eta = C\left(\dfrac{v - b}{v}\right) A$
1926	Duhief [34]	$\eta = \eta_0 \dfrac{v}{v - b}$
1933	Herzog and Kudar [35]	$\eta = \dfrac{1}{6}\sqrt{\dfrac{8RTM}{\pi}}\, \dfrac{d}{V - \beta}$
1939	van Wijk and Seeder [36] *	$\eta = \dfrac{C}{V(V - \beta)}\, e^{B(V,\,T)/T}$
1940	Bingham and Kinney [37]	$V - C = A\phi - B/\phi$
1941	Eyring [25]	$\eta = (2\pi RT)^{1/2}/M^{1/6} \cdot \rho \cdot (N^2 v_f)^{1/3} \cdot e^{E/RT}$ [From equation 32, p. 485, Ref. 25, by placing $V = M/\rho$ and free-space/molecule $= M v_f/N$]

* This is the same as van der Waals, Jr.'s formula, since $n \propto \dfrac{1}{v}$.

In the above formulas:

η = viscosity

ϕ = fluidity = $1/\eta$

ρ = density

v = specific volume = $1/\rho$; V = molar volume; v_f = free-space per gram = $v - b$

b = volume of one gram of molecules in close packing (van der Waals' b)

β = volume of one mole in close packing

n = number of molecules in unit volume

N = Avogadro's number
d = diameter of a molecule
m = mass of a molecule; M = molecular weight
k = Boltzmann constant
$A, B, C R$, = constants

It will be observed that all these investigators, with the exception of Eyring, related viscosity to reciprocal free-space. The principal variable in the above formula of Eyring is, of course, the exponential term which decreases as T increases. The exponential term is multiplied, however, by $T^{1/2}$ and $v_f^{1/3}$ which increase as T increases, and by ρ which decreases as T increases. The influence of free-space on viscosity is therefore somewhat obscured in this Eyring formula, which should perhaps be considered as an example of the dependence of viscosity on temperature rather than on free-space.

Viscosity as a Function of Molecular Weight in Homologous Series. The dependence of viscosity on molecular weight at a fixed temperature or at some reference temperature such as the melting or boiling point has been studied by a number of investigators. The following is a chronological record of the more important contributions to this phase of the subject.

Date	Investigator	Relationship
1909	Dunstan and Thole [38]	$\ln \eta = am + b$
1925	Macleod [39]	$\eta = Kma/x$
		where a = degree of association
		x = relative free-space
1931	Andrade [40]	$\eta_f = 2m\nu/\sigma$
		or $\eta_f = B(AT_f)^{1/2}/V_A^{2/3}$
		where ν = frequency of junction per second
		σ = average distance between centers
		A = atomic weight
		V_A = atomic volume
1938	Souders [41]	$\log_{10} \log_{10} \eta = I\rho/m - 2.9$
1939	Lewis and Morgan [42]	$\ln \eta = A(\ln m)/T + B \ln P + C$
		where P = parachor
1939	Hugel [24]	$\ln \eta - a = Q/R(T - b)$
		where a and Q are functions of m
1940	Flory [43]	$\ln \eta = am^{1/2} + b$
1940	Kauzmann and Eyring [44]	$\eta = Ae^{Evis/RT}$
		where E_{vis} is related to L, the latent heat of vaporization
1941	Linke [45]	$\ln \eta = A + B/4.57T$
		where A and B are functions of m
1943	Friend and Hargreaves [46]	$R = m\eta_b^{1/8}/\rho$

1944   Kierstead and Turkevich [47] $\phi = Ae^{-E/RT}$
                                        where A and E are functions of m

In the above formulas the symbols are:

  η = viscosity; η_f = viscosity at the melting point; η_b = viscosity at the boiling point

  ϕ = fluidity = $1/\eta$

  m = molecular weight

  T = absolute temperature; T_f = absolute temperature of fusion

  ρ = density

  a, b, A, B, C, I, K, R = constants

Of all formulas proposed expressing viscosity as a function of constitution, only the formula of Andrade above for the viscosity at the melting point enables the calculation of this value directly without the use of arbitrary constants. The agreement is good in the case of certain elements, but poor in many other instances.

The studies of Hugel, Kauzmann and Eyring, Linke, and Kierstead and Turkevich are similar in nature, since each is an attempt to define further the constants of the de Guzman-Andrade equation in terms of molecular weight. None of these relationships is applicable over an extended range of molecular weights.

The expression proposed by Souders is of considerable interest since it provides a means of estimating roughly the viscosity of a large number of substances when the density at the desired temperature is known. The viscosity-constitutional constant, I, can not only be calculated from atomic and structural constants, but has been shown by Lagemann [48] to be related to the refractive index and by Buehler [49] to the surface tension. A nomograph for organic liquids based on the Souders relation has been published by Davis.[50]

The proposal of Friend and Hargreaves to establish a "rheochor," R $m\eta_b^{1/8}/\rho$, resembling the parachor, $P = m\gamma^{1/4}/\rho$, does not seem to have attracted other investigators so far, although the idea seems to have interesting possibilities.

Viscosity as a Function of Concentration in Solutions. The empirical study of the relation between viscosity and concentration in solutions of resinous materials has been most extensive. Since this approach is not likely to yield information of fundamental significance unless the form of the empirical function is related to some physical concept of the mechanism of the flow process, little of value will be gained by attempting to enumerate the empirical formulas heretofore proposed. A few however deserve mention, because they indicate the type or form of such functions that have proved somewhat successful over limited ranges. The

following is a chronological record of some of the more important empirical functions proposed in the literature:

Date	Investigator	Relationship
1887	Arrhenius [51]	$\eta = A^{x_1} B^{x_2}$ For dilute solutions, this reduces to $\ln \eta_r = k \cdot c$
1913	Baker [52]	$\eta = \eta_0 (1 + ac)^k$
1932	Fikentscher [53]	$\ln \eta_r = \left[k + \dfrac{75k^2}{1 + 1.5kc} \right] c$
1935	Papkov [54]	$\ln \eta_r = kc^a$
1936	Philippoff [55]	$\eta_r = (1 + c[\eta]/8)^8$ (This is a special case of Baker's formula)
1937	Bredee and deBooys [56]	$\eta_r = \left[1 + \dfrac{2.5b}{6} c \right]^6$ (Also a special case of Baker's formula)
1942	Martin [57]	$\ln (\eta_{sp}/c) = \ln [\eta] + k[\eta]c$
1947	Doyle, Harbottle, Badger and Noyes [58]	$\eta_r = 1 + [\eta]c \cdot 10^{k[\eta]c}$ (Martin's formula)

In the above formulas:

η = viscosity

η_r = relative viscosity = $\dfrac{\text{viscosity of solution}}{\text{viscosity of solvent}}$

η_0 = viscosity of solvent

c = concentration, g/100 cc

x = volume fraction

$[\eta], A, B, a, b, k$ = constants

Of all the above relationships, the following four are the most widely used, and hence of greatest interest:

Andrade	$\eta = Ae^{B/T}$	(1)
Arrhenius	$\eta = Ae^{Bc}$	(2)
Dunstan	$\eta = Ae^{Bm}$	(3)
Flory	$\eta = Ae^{Bm^{1/2}}$	(4)

The Andrade function is of especial significance, firstly because of its remarkably good fit in most cases, and secondly because it serves as the basis of most of the attempts that have been made to interpret the mechanism of viscosity theoretically. The latter is undoubtedly so because of the closeness of fit provided by this function.

The Arrhenius function is the least general of the above four expressions, since it is usually limited to quite narrow ranges of concentration. Oddly enough, this function is often more applicable to non-Newtonian

solutions (such as concentrated "Vinylite" resin solutions), than it is to Newtonian solutions.

The Dunstan rule is likewise of quite limited application in homologous series of pure compounds, but is of interest when applied to solutions, as it forms the basis of the Staudinger rule for the determination of the molecular weights of polymers. This will be developed in a later section.
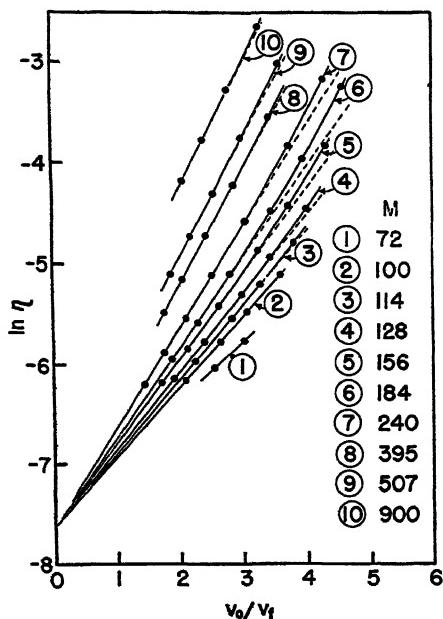
Figure 1. n-Paraffin hydrocarbons. Logarithm of viscosity vs. reciprocal relative free-space.

The Flory rule is of recent origin. It is generally useful for dealing with solutions in the intermediate to high concentration range, but cannot be applied successfully over the range from very low to intermediate concentrations.

The Free-Space Function

The four functions presented above are now known to be approximation functions rather than exact functions. The failure of these, and all other functions investigated by us, led the author to question whether either temperature or molecular weight was, in fact, a direct variable on which viscosity depends. It seemed more likely that both temperature and molecular weight influenced some other variable that more directly determined the viscosity. Such a variable was found to be the volume concentration of molecules in the liquid per unit volume of free-space.

This variable is influenced by both temperature and molecular weight (in homologous series). When used as the direct variable in a simple exponential equation, it reproduces the measured values with great accuracy except in the neighborhood of the freezing points. That is, the function.

$$\eta = A e^{B(v_0/v_f)} \tag{5}$$

(where v_0 = volume of molecules per gram of the liquid
v_f = free-space per gram of the liquid
A, B = constants for a single substance)

reproduces the measured viscosities with an accuracy equal to or better than the accuracy attainable in very high precision measurements. No other function examined by us approaches this degree of validity. We were therefore led to the belief that equation (5) may be an exact function.
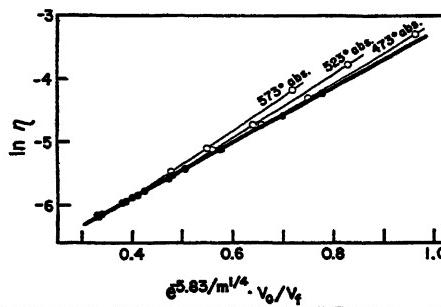
Figure 2. n-Paraffin hydrocarbons. Dependence of the logarithm of viscosity on reciprocal relative free-space when both temperature and molecular weight vary.

The work from which the above conclusions were drawn was a precision study of ten n-paraffin hydrocarbons ranging from pentane, $m=72$, to tetrahexacontane, $m=900$. The temperatures of measurement ranged from $-10°$ to $300°$C. When the logarithms of the viscosity were plotted (Figure 1) against the reciprocal relative free-space, straight lines resulted for each compound over all temperatures except those in the neighborhood of the freezing points. The lines for the paraffins from heptane to heptadecane inclusive converged to a common intercept, so it was possible to set up a molecular-weight function defining these values by determining the form of the dependence of their slopes on molecular weight. Thus it was found that:

$$B = C e^{-\beta'/m^{1/4}} \tag{6}$$

From this relationship, one can write:

$$\ln \eta = C e^{-\beta'/m^{1/4}} (v_0/v_f) + \ln A \tag{7}$$

Equations (6) and (7) are not exact functions in the sense that equation (5) is thought to be, since there is a slight divergence of the inter-

cepts from their average value. The agreement is close enough, however, for equation (7) to define all of the viscosities of paraffins $m = 100$ through 240 with an accuracy of about 1 per cent, except those values excluded because of proximity to their freezing points. The values for the higher-molecular-weight paraffins diverge from the base line at the higher temperatures. Figure 2 illustrates the application of equation (7) to the paraffin data from $m = 100$ through 900.

The Logarithmic-Decrement Function

Since the accurate determination of free-space is difficult for series of compounds other than the n-paraffins, it is desirable to set up a molecular-weight function that does not depend on free-space, even if some measure of general applicability is sacrificed by so doing. Accordingly, it was
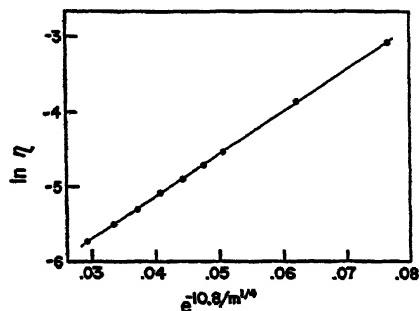
Figure 3. n-Acetic esters at 50°C. viscosities rectified according to logarithmic-decrement rule.

shown that, at constant temperature, v_0/v_f is approximately proportional to $e^{-K/m^{3/4}}$,

whence
$$B(v_0/v_f) \cong e^{-\beta'/m^{1/4} - K/m^{3/4}} \tag{8}$$

Now expressions showing a logarithmic-decrement type of variation, such as equation (7), afford considerable leeway in fitting data by adjustment of the constants. Thus, it was found that the variation of $K/m^{3/4}$ could be absorbed in $\beta'/m^{1/4}$ by making proper adjustment in each of the constants of the equation, resulting in the final form of the molecular-weight function at constant temperature of:

$$\ln \eta = \alpha e^{-\beta/m^{1/4}} + \gamma \tag{9}$$

On account of the adjustment made in the constants, however, the slope and intercept of equation (9) bear no relation to the corresponding slope and intercept of equation (7) and consequently lose their physical significance.

Equation (9) was successfully applied (as an approximation function) to the n-paraffins and to a variety of other homologous series of pure

compounds. An example is given in Figure 3. The fit of the data over an extended range of molecular weights was shown to be much better than that of either the Dunstan (linear) or the Flory (square-root) rules.

Homologous series of heterogeneous polymers may be treated by equation (9) without particular concern for the method of averaging used in determining their molecular weights. [If an average other than the number-average is used, the polymers should be sufficiently heterogeneous to make true the assumption that the particular average chosen is the same multiple of the number-average in all cases. This follows because the constant multiplier of the number-average is absorbed in the appropriate value of β. For example, if weight-averages are used, it is assumed that $\overline{M}_W = 2\overline{M}_N$, whence

$$\beta/\overline{M}_W^{1/4} = (1/2)^{1/4}\beta/\overline{M}_N^{1/4} = \beta''/\overline{M}_N^{1/4}]$$

An example of a homologous series of polymers rectified according to the logarithmic-decrement rule is given in Figure 4.
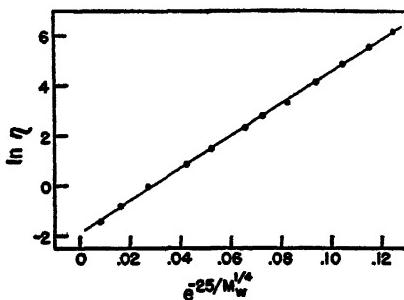
Figure 4. Decamethylene adipate polymers at 109°C. Data of Flory rectified according to logarithmic-decrement rule.

The Viscosity of Mixtures

The viscosity of mixtures of two (or more) pure compounds has been the subject of considerable study over many years. The early attempts to formulate expressions that would define the viscosity of mixtures were usually based on some method of combining the viscosities of the components. Most of these relationships may be shown to be variations of the following general form:

$$f(\eta) = x_1 f(\eta_1) + x_2 f(\eta_2) \tag{10}$$

in which x represents the concentration as weight fraction, volume fraction, mole fraction, etc., and $f(\eta)$ is some function of the viscosity. When $f(\eta)$ is η, we have the linear formula.[50] If $f(\eta) =$ Inη, the Arrhenius [51] formula results. With $f(\eta) = 1/\eta$, equation (10) becomes Bingham's [60]

formula, whereas if $f(\eta) = \eta^{1/3}$, the formula of Kendall and Monroe,[61] is represented. Kottler's [62] formula, which makes use of the Batschinski free-volume relation, may be thrown into the form of equation (10) if

$$f(\eta) = 1/\eta \quad \text{and if} \quad x_1 = \frac{N_1}{C_1{}^{N_1}C_2{}^{N_2}} \quad \text{and} \quad x_2 = \frac{N_2}{C_1{}^{N_1}C_2{}^{N_2}}$$

C_1 and C_2 being the Batschinski constants, and N_1 and N_2 being the mole fractions of components 1 and 2.

A different type formula was proposed by van der Wyk [63] by considering the interaction between the components of the mixture. His formula is:

$$\ln \eta = N_1{}^2 \ln \frac{\eta_1 \eta_2}{\eta_{12}{}^2} + 2N_1 \ln \frac{\eta_{12}}{\eta_2} + \ln \eta_2 \tag{11}$$

The only constant in this formula is η_{12}. It is claimed that equation (11) gives a better fit than the cube-root equation of Kendall and Monroe, which is probably the best of the formulas of the type of equation (10).

While equations of the type of (10) and (11) are useful in the case of two-component mixtures of normal liquids having molecular weights of the same order of magnitude, they fail as soon as the difference in molecular weights of the components becomes appreciable.

More recent approaches to the problem of the viscosity of mixtures have dealt with the molecular weights of the components rather than the viscosities of the components. Thus Flory,[43] Huggins,[64] and others have set up relationships defining viscosity as a function of certain average molecular weights of the components. The author's approach is along similar lines.

The Weighted-Average Molecular Weight. If one wishes to obtain the average of a group of numbers, he simply divides their sum by their number. If, however, he is concerned about the effect that a group of objects may have on some phenomenon that can be measured, it is possible that objects of different size or different mass may contribute different effects. Thus, when one speaks of average molecular weights of polymers, he is not dealing with abstract numbers but rather with the effect on measurable phenomena that the molecules produce. Consequently, the average molecular weight, $\overline{M}$, of a substance, containing n molecules $m_1, m_2, m_3 \ldots m_n$ is a weight such that n of these average molecules would produce the same effect as do the $m_1 + m_2 + m_3 + \ldots m_n$ actual molecules present. This means, generally speaking, that the nature of the phenomenon involved determines the kind of an average molecular weight one measures.

Thus, if the phenomenon is that of reacting the end groups with an appropriate reagent, it will make no difference whether the molecules are

large or small. Each will have two end groups and all one does by such a titration is to count their number. The use of this phenomenon as the criterion, therefore, is no different than dealing with abstract numbers. This average is the number-average which is written:

$$\overline{M}_N = \sum \left[\frac{n_1}{\Sigma n} m_1 + \frac{n_2}{\Sigma n} m_2 + \frac{n_3}{\Sigma n} m_3 + \cdots \right] = \frac{\Sigma nm}{\Sigma n} = \frac{\Sigma nm}{N} \qquad (12)$$

Suppose, however, that the phenomenon to be measured is influenced by the size or mass of the molecules, such as would be the case with diffusion or with internal pressure. In this case, the resulting effect would be determined by the distribution of molecular sizes present. It would therefore be necessary to weight the contribution of each size group by a factor involving the size of the members of that group. To be perfectly general, let such a weighting factor be $f(m) = m^a$. We should then write, for this general case: *

$$\overline{M}_a = \sum \left[\frac{n_1 m_1^a}{\Sigma n m^a} m_1 + \frac{n_2 m_2^a}{\Sigma n m^a} m_2 + \frac{n_3 m_3^a}{\Sigma n m^a} m_3 + \cdots \right] = \frac{\Sigma n m^{a+1}}{\Sigma n m^a} \qquad (13)$$

It is clear that if $a = 1$,

$$\overline{M}_{a=1} = \frac{\Sigma n m^2}{\Sigma n m} = \frac{\Sigma n m^2}{W},$$

and this particular average becomes the weight-average, $\overline{M}_W$. If $a = 0$,

$$\overline{M}_{a=0} = \frac{\Sigma n m}{\Sigma n} = \frac{\Sigma n m}{N},$$

which is the number-average.

For the case of a mixture of only two components whose molecular weights are m and M_a, and where w_2 is the weight fraction of M_a present, equation (13) reduces to:

$$\overline{M}_a = m + \frac{M_a - m}{1 + (M_a/m)^{1-a}(1 - w_2)/w_2} \qquad (14)$$

---

* In 1943, Flory[47] proposed the term "viscosity-average molecular weight" which he represented by the symbol, $\overline{M}_v$, defined by:

$$\overline{M}_v = \left[\frac{\Sigma n m^{a+1}}{\Sigma n m} \right]^{1/a}$$

In this equation, a is defined by

$$\frac{(\eta_{sp})_i}{c_i} = K m_i^a \qquad \text{for the } i\text{th polymer component,}$$

or by $[\eta] = K \overline{M}_v^a$ for the complete polymer.

Dr. Huggins has recently informed the author that the Committee on Nomenclature of the National Research Council proposes the symbol, $\overline{M}_a$, in place of Flory's $\overline{M}_v$, above. The symbol, $\overline{M}_a$, proposed by the Committee has a different meaning from that used by the author in equation (13), although both reduce to the same expression when $a = 1$ (i.e., for a weight-average molecular weight). No change could be introduced in this text because the plates had already been made when the printer was approached regarding it.

The viscosity of mixtures at constant temperature may be readily rectified by the logarithmic-decrement function, equation (9), if one uses the appropriate weighted-average molecular weight of the components of the mixture, averaged according to equations (13) or (14). It was found by examination of a number of systems, however, that there is no
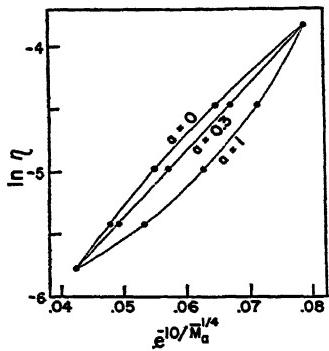
Figure 5. n-Heptane and n-heptadecane at 50°C. Viscosities of mixtures rectified according to logarithmic-decrement rule, showing influence of parameter, a.

definite "viscosity-average" molecular weight such as we speak of in the case of the number-average or the weight-average, but the value of the parameter, a, must be determined for each system studied.

A number of examples of the viscosities of mixtures rectified according to the logarithmic-decrement equation are given in Figures 5, 6, 7, and 8. It will be noted that the value of the parameter, a, varies from 0.3 to 1 in this particular group of mixtures.
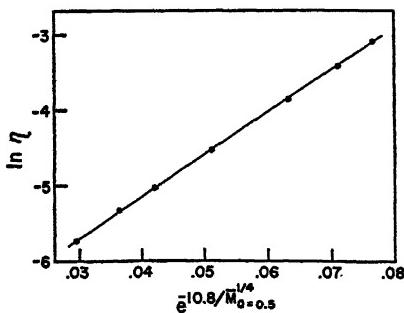
Figure 6. Ethyl acetate and n-octadecyl acetate at 50°C. Viscosities of mixtures rectified according to logarithmic-decrement rule.

The Viscosity of Solutions of Resinous Substances

The principal contribution to polymer chemistry that is made by the research work here described lies in the rectification of the dependence of viscosity in Newtonian flow on concentration in polymer solutions. More study has probably been given to this aspect of Newtonian flow in recent

years than to any other problems involving viscosity, because of its bearing on the manufacture and use of resinous substances, as mentioned at the beginning of this chapter.

The field is rather sharply divided between studies made on dilute solutions and those on concentrated solutions, the latter extending to
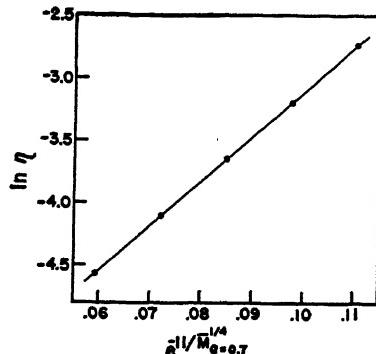
Figure 7. Di-*n*-butyl and di-*n*-octadecyl succinate at 100°C. Viscosities of mixtures rectified according to logarithmic-decrement rule.

the undiluted molten polymers. No function has heretofore been adequate to define accurately the dependence of viscosity in Newtonian flow on concentration (or on average molecular weight) from very dilute solutions clear through to the undiluted polymer. Consequently, the two divisions of the field have been investigated separately, by far the most emphasis being placed on the dilute-solution category.
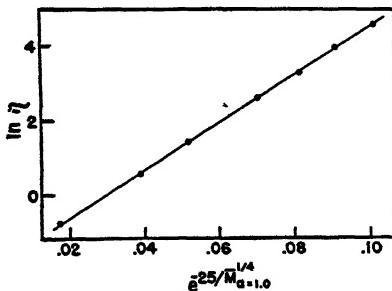
Figure 8. Decamethylene adipate polymers, *m* equals 1402 and *M* equals 14,110, at 109°C. Data of Flory rectified according to logarithmic-decrement rule.

Investigation of dilute solutions of polymers has yielded two important results of great value in polymer chemistry. The first of these is the application of hydrodynamic principles to the theory of the viscosity of dilute solutions of macromolecules.

Several excellent reviews have been published of the work done by investigators following the hydrodynamic type of approach to the study of the flow of dilute solutions of macromolecular substances. One in-

terested in this field should certainly read the Burgers,[65] Lauffer,[66] and Eirich [67] reports. The subject is also reviewed in a forthright manner in Mark's [68] book. It will suffice here to say that the students of the hydro-dynamic approach have extended the classical treatment of Einstein [69] on the viscosity of dilute suspensions of rigid spheres to the case of solu-tions of rod-shaped macromolecules, under circumstances where Brownian movement either may or may not be a factor. Expressions for the specific viscosity in terms of the concentration, c, and of the axial ratio, f, have been derived, that have been substantiated in certain cases by experi-ments made with model suspensions such as chopped silk fibers in water. These studies have contributed materially to our knowledge of the size and shape of macromolecules in solution.

Another achievement of great value in polymer chemistry, resulting from the study of the viscosity of dilute solutions, is the observation by Staudinger that the molecular weight of the dissolved polymer is propor-tional to the specific viscosity divided by the concentration, or, as more often written,

$$\eta_{sp} = K_m M_w c \tag{15}$$

$$\text{where} \quad \eta_{sp} = \eta_r - 1 = \frac{\eta \text{ soln.}}{\eta \text{ solv.}} - 1$$

M_w = weight-average molecular weight of the dissolved polymer.

c = concentration of solute, g/100 cc solution

K_m = "Staudinger constant"

This relationship may be deduced from the Dunstan rule, equation (3), in the following manner:

$$\ln \eta = B\overline{M}_w + \ln A \tag{3}$$

Where $\overline{M}_w$ = weight-average molecular weight of solvent plus polymer.

But [see equation (20)] $\overline{M}_w = m + (M_w - m)w_2$

$$= m + (M_w - m)(c/100) \cong m + M_w(c/100)$$

Therefore $\ln \eta = BM_w(c/100) + (Bm + \ln A) \tag{16}$

When $(Bm + \ln A) = \ln \eta_{solv.}$ and $B = 100 \, K_m$

equation (16) becomes $\ln \eta_r \cong \eta_{sp} = K_m M_w c$

The simple Staudinger rule has not been found adequate to apply to all systems of polymer plus solvent, and it has accordingly been modified in several ways to provide a better fit. One of the most useful of these modifications is simply the addition of a c^2 term.

Thus, if $\ln \eta_r = K_m M_w c + K_2 c^2 \tag{17}$

Then, $\dfrac{\ln \eta_r}{c} = K_m M_w + K_2 c \tag{18}$

By plotting $\dfrac{\ln \eta_r}{c}$ versus c, a straight line is usually obtained whose inter-
cept is $K_m M_w$. The value of the left-hand member of equation (18) at
$c=0$ was called the "intrinsic viscosity," $[\eta]$, by Kraemer,[70] and this term
has become very useful in polymer chemistry.

The viscosity of concentrated solutions of polymers has received far less
attention by investigators than the viscosity of dilute solutions. This is, in
part, owing to the departure from Newtonian conditions observed in the
case of many concentrated polymer solutions.

A few examples of empirical functions relating the viscosity of solu-
tions to concentration were given in an earlier section. None of these
examples is of general application. Each applies to a particular class of
polymers and each, in general, is valid over restricted ranges of concen-
tration. The Arrhenius rule is most convenient when used in the form
$\ln \eta_r = kc$. This relationship has been applied to shellac solutions by Ver-
man [71] and to "Vinylite" solutions in numerous solvents in our own lab-
oratories. The Papkov formula is a logical extension of the Arrhenius
rule, and has been applied to solutions of phenolic resins in acetone by
Klaassens and Houwink.[72] Baker's formula perhaps has the most uni-
versal application of any of those listed. It has been used in various forms
by numerous investigators, particularly Philippoff,[55] Bredee and de-
Booys, [56, 73] and Mead and Fuoss.[74] One of the more recent concentration
formulas was proposed by Martin [57] in 1942. This was applied with some
success to solutions of polystyrene by Spencer and Williams,[75] and to
nitrocellulose solutions by Doyle, Harbottle, Badger, and Noyes.[58]

Considerable encouragement to the feeling that a true viscosity func-
tion should be possible of attainment was given by the papers of Irany,[76]
published in 1938 to 1941, in which he showed graphically that the de-
pendence of viscosity on concentration (as well as on temperature) could
be rectified by the use of a suitable scale. The nature of Irany's function
has not, however, been elucidated.

The most useful relationship heretofore proposed for dealing with the
dependence of viscosity on concentration in the intermediate- to high-
solids range is that of Flory,[43] who first proposed to deal with solutions
as mixtures of members of homologous series in which the variable is
the weighted-average molecular weight of solvent and solute. Flory's well-
known square-root rule,

$$\ln \eta = \ln A + B \bar{M}_w^{1/2} \qquad (19)$$

has been applied to solutions of decamethylene adipate polymers in
diethyl succinate by Flory [77] himself, to solutions of polystyrene in var-
ious solvents by Spencer and Williams,[75] and to other systems by a num-
ber of investigators. The author's approach to this subject may be con-

sidered as an extension of the Flory treatment, although our work leading to this concept was done independently some years prior to Flory's.

Adaptation of the Logarithmic-Decrement Equation to the Viscosity of Solutions of High Polymers. The treatment of solutions of high polymers, proposed by the author, involves a slight extension and some simplification of the ideas advanced in the prior section on mixtures. The extension arises from the necessity of considering the solvent as a member of the same homologous series as the resinous component or solute. In most cases this offers no serious complications, but in the case of solvent and solute of quite dissimilar types the significance of the weighted-average molecular-weight parameter is lost. The simplification arises from the fact that, in solutions of high polymers, the spread in molecular weight between solvent and solute is so great that such solutions may be treated as special cases of the more general rules that apply to mixtures. Thus, if solvent and solute do belong to the same homologous series, the great spread between their molecular weights usually calls for a value of *one* for the weighted-average molecular-weight parameter, *a,* of equation (14). Under such circumstances equation (14) reduces at once to:

$$\overline{M}_{a\,=\,1} = \overline{M}_w = m + (M_w - m) \cdot w_2 \qquad (20)$$

If, however, solvent and solute are of quite different chemical families, a value of *a* somewhat less than *one* often has been found to apply. If this value is 0.5 or less, a different method of simplification of equation (14) may be used. Thus, we may write:

$$\overline{M}_a = m + \frac{M_a - m}{1 + (M_a/m)^{1-a}(1 - w_2)/w_2}$$

$$= m + \frac{M_a - m}{(M_a/m)^{1-a}} \cdot \frac{w_2}{1 - w_2[1 - 1/(M_a/m)^{1-a}]} \qquad (21)$$

If M/m is a sufficiently large number, $1/(M_a/m)^{1-a}$ becomes negligible in comparison with unity for values of *a* equal to or less than about 0.5. This simplification is valid for low values of w_2 but differs appreciably from the true function as w_2 approaches unity. The simplified expression is:

$$\overline{M}_{\underline{a}\,\leq\,0.5} = m + \frac{M_a - m}{(M_a/m)^{1-a}} \frac{w_2}{1 - w_2} \qquad (22)$$

Dependence on Molecular Weight. If the molecular weight of the polymer is known, equations (20) or (22)—whichever is appropriate to the system studied—may be used as written. Figure 9 presents Flory's data on solutions of decamethylene adipate polymer, $M_w = 16,070$, in diethyl succinate, $m = 174.2$, at 79°C, rectified according to the logarith-

mic-decrement rule using the simplification indicated in equation (20). Attention is particularly directed to the good conformance of the data at very low concentrations of solute. The same data plotted according to the square-root rule, Figure 10, fail to conform in this region.
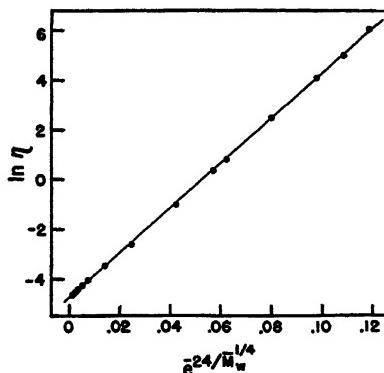
Figure 9. Decamethylene adipate polymer in diethyl succinate at 79°C. Data of Flory rectified according to logarithmic-decrement rule.

Dependence on Concentration. In most industrial applications where viscosity control or blending operations are carried out, the molecular weight is not known. In such cases it is desirable to rectify viscosity as a function of solute concentration without reference to molecular weight. Equation (9) is readily modified for this purpose by further simplification
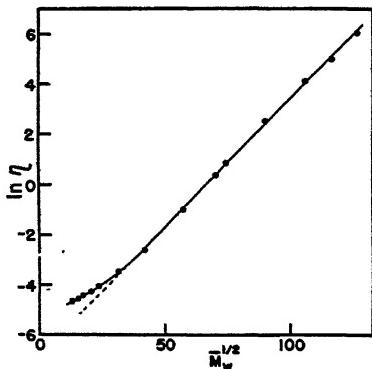
Figure 10. Decamethylene adipate polymer in diethyl succinate at 79°C. Data of Flory rectified according to square-root rule.

of equations (20) and (22). Before proceeding with this step, however, it is necessary to consider certain ways of expressing concentration in solutions. In industrial work, concentration is almost always expressed as per cent of solids or grams of solute per 100 grams of solution. In scientific work, however, it is often useful to express concentration in grams of solute per 100 grams of solvent. The symbol c is used in either case, with

the appropriate units indicated. Now, for any given system of solvent and solute, both the $(M_w - m)$ of equation (20) and the $(M_a - m)/(M_a/m)^{1-a}$ of equation (22) are fixed. Likewise, $w_2 = c/100$ grams of solute per gram of solution, and $w_2/(1 - w_2) = c/100$ grams of solute per gram of solvent. Consequently, both equations (20) and (22) may be written:

$$\overline{M}_a = m + kc \qquad (23)$$

for the particular values $a = 1$ and $a \leq 0.5$, provided the unit of concentration is properly defined.

For the special cases just discussed, equation (9) could be written:

$$\ln \eta = \alpha e^{-\beta/(m + kc)^{1/4}} + \gamma \qquad (24)$$

but a further simplification is desirable and may be made by translating the axes defining $\ln \eta = f(\overline{M}_a)$ a distance of m units to the right. The same curve is then defined with respect to the new position of the axes by $\ln \eta = f(kc)$, since $\overline{M}_a = m + kc$. Therefore, for high polymer solutions we may write,

$$\ln \eta = \alpha e^{-\beta/(kc)^{1/4}} + \gamma = \alpha e^{-K/c^{1/4}} + \gamma \qquad (25)$$
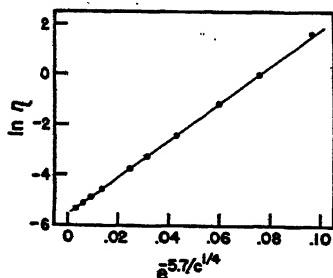
Figure 11. Dependence on concentration. Half-second nitrocellulose in acetone at 20°C. Viscosities of high-polymer solutions rectified according to logarithmic-decrement rule. (Concentration in grams of solute per 100 grams of solvent.)

for the special cases where the weighted-average molecular-weight parameter, a, either is unity or else is 0.5 or less. In the former case, c is in grams of solute per hundred grams of solution, α is the same as in equation (9), and γ is $\ln \eta_{solv}$. In the latter case, however, an approximation is made that introduces an appreciable deviation toward the higher concentration end. In order to rectify the viscosity as a function of concentration in this case, a slight adjustment to take care of this deviation must be made in the β of equation (9), requiring somewhat different values of α and γ in equation (25).

Equation (25) is of wide applicability to polymer solutions that exhibit Newtonian flow. Figure 11 illustrates the use of equation (25) with data on the viscosities of acetone solutions of nitrocellulose at 20°C. In this case, the value of the parameter happens to be 0.5 so that concentration

is expressed in grams of solute per 100 grams of solvent. The agreement is good from 1 to 20 per cent solids (1 to 25 grams per 100 grams of solvent).

Better agreement over a broader range of concentrations is obtained by the use of equation (9) which involves the molecular weight of the solute. Figure 12 illustrates this case with the same solutions of nitrocellulose in acetone that were used in the previous illustration. For these calculations, a value of $M_{a=0.5} = 23,250$, obtained by multiplying the number-average by three-halves, was used. The number-average molecular weight of the nitrocellulose was determined to be 15,500 by the osmotic pressure method. The agreement obtained in this study by the use of equation (9) is considerably better than that shown by the use of (25),

Figure 12. Dependence on molecular weight. Half-second nitrocellulose in acetone at 20°C. Viscosities of high-polymer solutions rectified according to (1) linear rule, (2) square-root rule, (3) logarithmic-decrement rule.
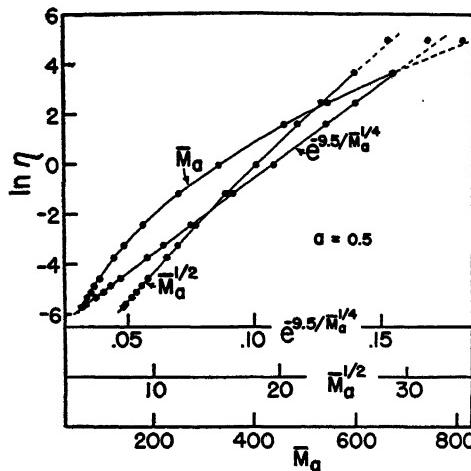
since values from zero to 35 per cent solids are rectified by the former, whereas only values from 1 to 20 per cent solids are rectified by the latter equation. Comparison is also made in Figure 12 of the logarithmic-decrement rule [equation (9)] with the linear and with the square-root rules. The greater range of usefulness of the former rule is quite apparent in this illustration. Of particular interest is the fact that the logarithmic-decrement rule, equation (9), rectifies the data at very low concentrations as well as through the intermediate to higher concentrations. In spite of this excellent conformity, it must be remembered that the logarithmic-decrement rule is not an exact function. It is only a good approximation and its use to extrapolate values beyond the limits of the data should not be attempted. Thus, in the case of the study of acetone solutions of nitrocellulose referred to above, deviations from the straight

line of Figure 12 are considerable at 40 per cent solids and become as great as 18.5 per cent at 50 per cent solids.

REFERENCES

1. Poiseuille, J. L. M., *Mem. Savants Etrangers*, **9**, 433 (1846).
2. Slotte, K. F., *Beibl. Ann. Physik*, **16**, 182 (1892).
3. Porter, A. W., *Phil. Mag.*, Ser. 6, **23**, 458 (1912).
4. Nissan, A. H., Clark, L. V. W., and Nash, A. W., *J. Inst. Petroleum*, **26**, 155 (1940).
5. Telang, M. S., *J. Phys. Chem.*, **50**, 373 (1946).
6. Bingham, E. C., *Am. Chem. J.*, **43**, 287 (1910).
7. ———, and Stookey, S. D., *J. Am. Chem. Soc.*, **61**, 1625 (1939).
8. de Guzman, J., *Anales soc. españ. fis. quim.*, **11**, 353 (1913).
9. Arrhenius, S., *Medd. Vetenskapsakad. Nobelinst.*, **3**, (20), 1-40 (1916).
10. Kendall, J., and Monroe, K. P., *J. Am. Chem. Soc.*, **39**, 1787 (1917).
11. Drucker, C. *Z. physik. Chem.*, **92**, 287 (1918).
12. Vogel, H., *Physik. Z.*, **22**, 645 (1921).
13. Raman, C. V., *Nature*, 111, 532 (1923).
14. Frenkel, J., *Z. Physik*, **35**, 652 (1926).
15. Dunn, J. S., *Trans. Faraday Soc.*, **22**, 401 (1926).
16. Andrade, E. N. da C., *Nature*, **125**, 309 (1930).
17. Madge, E. W., *J. Phys. Chem.*, **34**, 1599 (1930).
18. Sheppard, S. E., and Houck, R. C., *J. Rheol.*, **1**, 349 (1930).
19. Silverman, D., *Trans. Faraday Soc.*, **29**, 1285 (1933).
20. Cragoe, C. S., *Proc. World Petroleum Congr., London*, **2**, 529 (1933).
21. Macleod, D. B., *Trans. Faraday Soc.*, **32**, 872 (1936).
22. Eyring, H., *J. Chem. Phys.*, **4**, 283 (1936).
23. Barr, G. A., *Ind. Eng. Chem.*, **16**, 789 (1937).
24. Hugel, G., *Oel Kohle Erdoel Teer*, **15**, 27 (1939).
25. Glasstone, S., Laidler, K. J., and Eyring, H., "Theory of Rate Processes," p. 486, New York, McGraw-Hill Book Co., Inc., 1941.
26. Stuart, A. H., *Petroleum*, **10**, 74 (1947).
27. Walther, C., *Erdöl u. Teer*, **4**, 510, 526, 614 (1928); **5**, 619 (1929).
28. ———, *Petroleum*, **26**, 755 (1930).
29. ———, *Erdöl u. Teer*, **7**, 382 (1931).
30. ———, *Proc. World Petroleum Congr., London*, **2**, 419 (1933).
31. Batschinski, A. J., *Z. physik. Chem.*, **84**, 644 (1913).
32. van der Waals, J. D., Jr., *Proc. Acad. Sci. Amsterdam*, **21**, 743 (1918).
33. Macleod, D. B., *Trans. Faraday Soc.*, **19**, 6 (1923).
34. Dubief, J., *Compt. rend.*, **182**, 688 (1926).
35. Jaeger, F. M., "Second Report on Viscosity and Plasticity," p. 53, New York, Nordeman Publishing Co., Inc., 1938.
36. van Wijk, W. R., and Seeder, W. A., *Physica*, **6**, 129 (1939); **7**, 45 (1940).
37. Bingham, E. C., and Kinney, P. W., *J. Applied Phys.*, **11**, 192 (1940).
38. Dunstan, A. E., and Thole, F. B. *J. chim. phys.*, **7**, 204 (1909).
39. Macleod, D. B., *Trans. Faraday Soc.*, **21**, 151 (1925).
40. Andrade, E. N. da C., *Nature*, **128**, 835 (1931).
41. Souders, M., *J. Am. Chem. Soc.*, **60**, 154 (1938).
42. Lewis, D. T., and Morgan, A. R., *J. Chem. Soc.*, 1341, (1939).
43. Flory, P. J., *J. Am. Chem. Soc.*, **62**, 1057 (1940); **65**, 372 (1943).
44. Kauzmann, W. J., and Eyring, H., *J. Am. Chem. Soc.*, **62**, 3113 (1940).
45. Linke, R., *Z. physik. Chem.*, **188A**, 11 (1941).
46. Friend, J. N., and Hargreaves, W. D., *Phil. Mag*, **34**, 643 (1943).
47. Kierstead, H. A., and Turkevich, J., *J. Chem. Phys.*, **12**, 24 (1944).
48. Lagemann, R. T., *J. Am. Chem. Soc.*, **67**, 498 (1945).
49. Buehler, C. A., *J. Phys. Chem.*, **42**, 1207 (1938).

50. Davis, D. S., *Ind. Eng. Chem.*, **33**, 1537 (1941).
51. Arrhenius, S., *Z. physik. Chem.*, **1**, 285 (1887).
52. Baker, F., *J. Chem. Soc.*, **103**, 1653 (1913).
53. Fikentscher, H., *Cellulosechemie*, **13**, 58 (1932).
54. Papkov, S., *Kunststoffe*, **25**, 253 (1935).
55. Philippoff, W,. *Cellulosechemie*, **17**, 57 (1936).
56. Bredee, H. L., and de Booys, J., *Kolloid-Z.*, **79**, 31 (1937).
57. Martin, A. F., in "Cellulose and Cellulose Derivatives," ed. E. Ott, p. 966, New York, Interscience Publishers, Inc., 1943.
58. Doyle, G. J., Harbottle, G. Badger, R. M., and Noyes, R. M., *J. Phys. Coll. Chem.*, **51**, 569 (1947).
59. Dunstan, A. E., and Thole, B. T., "The Viscosity of Liquids," p. 41, London, Longmans, Green & Co., Ltd., 1914.
60. Bingham, E. C., *Am. Chem. J.*, **35**, 195 (1906).
61. Kendall, J., and Monroe, K. P., *J. Am. Chem. Soc.*, **39**, 1802 (1917).
62. Kottler, F., *J. Phys. Chem.*, **47**, 277 (1943); **48**, 76 (1944).
63. van der Wyk, A. J. A., *Nature*, **138**, 845 (1936).
64. Huggins, M. L., *Ind. Eng. Chem.*, **35**, 980 (1943).
65. Burgers, J. M., "Second Report on Viscosity and Plasticity," p. 113, New York, Nordemann Publishing Co., Inc., 1938.
66. Lauffer, M. A., *Chem. Revs.*, **31**, 561 (1942).
67. Eirich, F., *Phys. Soc. Repts., Progress in Physics*, **7**, 329 (1940).
68. Mark, H., "Physical Chemistry of High Polymeric Systems," p. 227, New York, Interscience Publishers, Inc., 1940.
69. Einstein, A., *Ann. Physik*, **19** (4), 289 (1906).
70. Kraemer, E. D., and Van Natta, F. J., *J. Phys. Chem.*, **36**, 3175 (1932).
71. Verman, L. C., *London Shellac Research Bur., Tech. Paper 11* (1937).
72. Klaassens, K. H., and Houwink, V. R.. *Kolloid-Z.*, **76**, 217 (1936); **99**, 160 (1942).
73. Bredee, H. L., and de Booys, J., *Kolloid-Z.*, **91**, 39 (1940); **99**, 171 (1942).
74. Mead, D. J., and Fuoss, R. M., *J. Am. Chem. Soc.*, **64**, 277 (1942).
75. Spencer, R. S., and Williams, J. L, *J Colloid Sci.*, **2**, 117 (1947).
76. Irany, E. P., *J. Am. Chem. Soc.*, **60**, 2106 (1938); **61**, 1734 (1939); **63**, 2611 (1941).
77. Flory, P. J., *J. Phys. Chem.*, **46**, 872 (1942).

THE IMPORTANCE OF COLLOIDS FOR CATALYSTS AND TRACE ELEMENTS

Alfons Krause

Inst. for Inorganic Chemistry, University of Poznań, Poland

CATALYSTS that are active in a heterogeneous system are often colloidal substances. The dispersion of solid substances or a cold working (*Kaltbearbeitung*) of them as, for example, the damaging of their surface by chipping, gives them specific structural properties. In these conditions the space positions of several atoms or ions undergo a change, and as a result the distance between them is not normal everywhere. The consequence is that their outermost electronic layer undergoes important changes, so that partially there can be formed free radicals or radical-like structures.[1] The degree of freedom of the radicals in a solid is small and therefore their life-period is very long.

Free Radicals in Catalysts

Free radicals, which can be dealt with as active centers [2] on the surface of the catalyst, enter easily into chemical reaction with other molecules because their outermost electronic layer is incomplete. They can exist in all active transitory phases that possess space-lattices with certain defects or disturbances. To the same type of phases belong ferromagnetic * compounds,[1, 3] especially amorphous substances, because their molecules can exist .in a state of a particular disorganization. The existence of radicals in active transitory phases [4] is affirmed by the following facts: (1) In different compounds, as e.g., in hydroxides and oxides of metals, especially those of heavy metals, the existence of the covalent (homopolar) bond should be accepted.[5] (2) Transitory phases, that is substances with space-lattice disturbances (*Gitterstörung*) have abundant energy.[6] (3) Transitory phases often show a considerable increase of paramagnetism.[7] The magnetic sensitiveness (susceptibility) diminishes with the

* See paper by L. J. E. Hofer, page 113 of this volume.

formation of normal space-lattice and disappearance of radical structure, e.g., at higher temperatures.

Thus, strongly heated, well-crystallized α-Fe$_2$O$_3$ (dark colored with a violet shade) has no or very slight catalyst activity, whereas colloidally dispersed (radical) ferric oxide (red) possesses catalytic properties and decomposes hydrogen peroxide, for instance, just as the gels of amphoteric metallic hydroxides, though not so strongly. The fragmentary structure of α-Fe$_2$O$_3$ can be represented as follows:

$$O=Fe—O—Fe=O \qquad\qquad O=Fe—O—\quad\quad—Fe=O$$

Stable ferric oxide   *Radical ferric oxide*

[well crystallized; not active]   [colloidally dispersed (transitory phase); good catalyst]

The radical $O=Fe—$ activates hydrogen peroxide, and as a result, the radical $\overset{\nearrow}{H}O$ appears and forms an uninterrupted chain of catalytic reactions:

$$O=Fe— + H_2O_2 \longrightarrow O=Fe—OH + \overset{\nearrow}{H}O$$
$$\overset{\nearrow}{H}O + H_2O_2 \longrightarrow H_2O + \overset{\nearrow}{H}O_2$$
$$\overset{\nearrow}{H}O_2 + H_2O_2 \longrightarrow H_2O + O_2 + \overset{\nearrow}{H}O$$
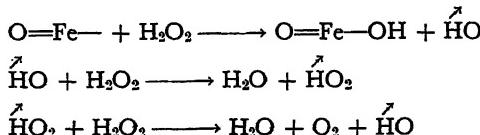
Following this conception, a series of other catalytic reactions have been elaborated, among others the synthesis of methanol and the method of Fischer-Tropsch.[1] Generally speaking, the catalyst, K, which has a suitable radical structure, causes the orientation and deformation of the molecules, A and B (which have to react) and changes them into radicals ($\nearrow$) or radical-like structures; this makes the chemical reaction between them possible.[1]

$$\overset{\nearrow}{K} + A \longrightarrow \overset{\nearrow\nearrow}{KA} + B \longrightarrow \overset{\nearrow\nearrow\nearrow}{KAB} \longrightarrow \overset{\nearrow}{K}[\overset{\nearrow}{A}\overset{\nwarrow}{B}] \longrightarrow \overset{\nearrow}{K} + [AB]$$
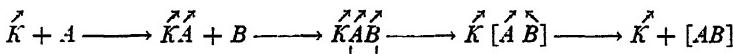
The reaction is accompanied by a genuine electronic resonance which regulates the catalytic phenomena as in a pendulum-like motion. The reaction is interrupted if the radical catalyst forms a stable compound with a certain substance acting as a "poison," which blocks up the active centers (radicals) of the catalyst.

The Composite Catalyst Surface

An important influence on the activity of catalysts is exerted by activators or promotors which are already active in minimal quantities. In 1939 we stated that cupric hydroxide in a quantity corresponding to 0.001 mg Cu strongly accelerates the peroxidatic activity of orthoferric

hydroxide in oxidization of formic acid by means of hydrogen peroxide at 37°C.[8] In this case there appears a complicated oxidation-reduction system acting on the basis of multi-levelled step (graduated) catalysis. Such a system possesses a relatively small energy of activation and is similar to the natural respiratory system, where instead of hydrogen peroxide, oxygen is the final acceptor.[9] Traces of cupric hydroxide act still more distinctly in the reaction of peroxidatic oxidation (decoloration) of the juice of red beets. Even as little as 10^{-9} mg Cu together with ferric hydroxide accelerate this reaction at 37°C, in such dilution as 1:50 billions.[10] The traces of copper, mentioned above, do not act by themselves but only when accompanied by ferric hydroxide, which serves as a suitable colloidal carrier. Mixed inorganic catalysts, having a suitable structure, are not inferior in their activity to natural ferments. In fact, there is an artificial inorganic ferment surpassing natural catalase in activity. Iron present in this ferment is especially active. As little as 10^{-7} mg Fe (as ferric hydroxide) has a positive influence on the decomposition of hydrogen peroxide at 37°C, under the condition that the trace of ferric hydroxide is accompanied by a suitable carrier composed of the gel of cupric and cobaltous hydroxides.[10] The above-mentioned quantity of iron is found in 1 mg of the superferment and corresponds approximately to the amount of iron in natural live tissue. A structure similar to that of the natural ferments is also found in other inorganic catalysts composed of the gels of various amphoteric metallic hydroxides and of ions of various metals adsorbed on them. These ions should be looked upon as a prosthetic group and the gels as carriers or apoferments.

This combination acts well if the several components are present in a well chosen unit. The carrier or ions by themselves act very slightly or do not show any catalytic properties at all; but together they can form very good catalysts, possessing specific and selective properties which can influence the direction of the catalytic reaction.

The Mn^{++} ion (10^{-6} g) on the carrier $Mg(OH)_2$ (gel) accelerates the catalytic decomposition of hydrogen peroxide at 37°C, but checks the peroxidatic oxidation (decoloration) of indigocarmine. The Cu^{++} ion behaves reversely. The influence of trace elements often depends upon their concentration, and sets of ions can cause a positive superadditive activity in some cases, but in other cases there may be a distinct ionic antagonism. The Sr^{++} ions (10^{-2} g) on the carrier $Bi(OH)_3$ (gel) check the catalytic decomposition of hydrogen peroxide slightly, but in the concentration 10^{-4} g they accelerate this reaction. The most advantageous in this case is the ionic group: Sr^{++} (10^{-7} g)$+Cu^{++}$ (10^{-7} g). Even the Na^+ ions act catalytically, but only if they are on the carrier $Zn(OH)_2$ (gel) and accompanied by the ions Cu^{++} and Co^{++}. The Co^{++} ions (10^{-6} g) on the carrier $Al(OH)_3$ (gel), which accelerate strongly the peroxidatic oxidization of

indigocarmine, are paralyzed in their activity by the UO_2^{++} ions (10^{-6} g). The above-quoted examples show that the sociology of chemical elements is of great importance for catalytic phenomena.

The inorganic ferments of the type of catalase and peroxidase are very common among inorganic compounds (e.g. amphoteric hydroxides) and useful materials (glass, cement, bricks, coal, chemical fertilizers, and even minerals).[11] Some of them can facilitate the oxidation of pyrogallol, benzidine, various dyes and many organic acids (even acetic acid) at room temperature.[9, 10] The benzidine reaction is also obtained by various metals, e.g. metallic platinum (best if roughened), which causes this reaction [in a slightly acidic medium ($0.1N$ CH_3COOH)] almost as well as blood. The reaction can be "poisoned" by traces of As_2O_3 (0.02 mg) or KCN (0.003 mg) and NaF (28 mg). Still more sensitive to "poisons" in the benzidine reaction is palladium—0.00002 mg KCN are sufficient.[12] But sometimes traces of "poisons" can improve the catalytic ability of the catalyst. A certain colloidal mixed catalyst composed of γ-FeOOH + CuO (0.2250 g) acts better in the peroxidatic oxidation of formic acid after the addition of $2.5.10^{-5}$ mg As_2O_3; a greater amount of As_2O_3 (25 mg) paralyzes its activity.[13]

There are also inorganic oxidases transporting molecular oxygen on to several organic and inorganic compounds.[10] The gel of cobaltous hydroxide (10^{-6} g) is a good catalyst in the oxidization of indefinite quantities of Na_2SO_3 with oxygen of the air at 20°C. The Co^{++} ion is not active in this case. The advantageous activity of $Co(OH)_2$ can be greatly increased in an alkaline medium. Probably the activity of radical-like hydrogen atoms which are present in the active OH-groups of cobaltous hydroxide can be greatly increased under the influence of OH- ions. As a result, the molecule of oxygen drawn into the effective sphere of active hydrogen atoms becomes active too. In this case free radicals can appear:

$$\text{Co}\begin{array}{c} \diagup O—H \\ \diagdown O—H \end{array} + 2OH^- \longrightarrow \text{Co}\begin{array}{c} \diagup O \ldots H \ldots OH^- \\ \diagdown O \ldots H \ldots OH^- \end{array}$$

$$\text{Co}\begin{array}{c} \diagup O \ldots H \\ \diagdown O \ldots H \end{array} + O_2 \rightarrow \text{Co}\begin{array}{c} \diagup O \ldots H \\ \diagdown O \ldots H \end{array}OO + 2Na_2SO_3 \rightarrow 2Na_2SO_4 + \text{Co}\begin{array}{c} \diagup O \ldots H \\ \diagdown O \ldots H \end{array}$$
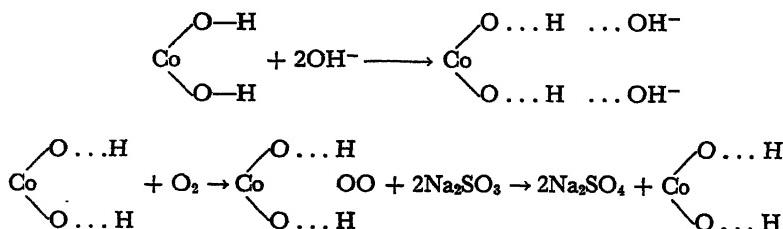
The gel of $Ni(OH)_2$ acts like $Co(OH)_2$. When in its presence the oxidization of sodium sulfite becomes complete, $Ni(OH)_2$ also becomes oxidized because of the inductive action of Na_2SO_3, and a black compound of trivalent nickel appears. Under the influence of a surplus of Na_2SO_3 (and also of H_2O_2), the black compound again changes into the green $Ni(OH)_2$.[10]

The catalytic oxidation of arsenic with oxygen of the air can be performed at pH 10 in the presence of Cu(OH)$_2$ gel as catalyst. Cupric hydroxide can be greatly activated by the addition of other amphoteric hydroxides.[10] The gel of orthoferric hydroxide by itself, or in combination with various activators, does not act catalytically in the above-mentioned reaction. Most probably its active hydrogen atoms do not possess a sufficient degree of activation as in the case of cupric hydroxide. Ferric hydroxide alone strongly sorbs As$_2$O$_3$, which blocks the radical centers on its surface, forming stable complex compounds. The Fe(OH)$_2$ gel is a good inductor in this reaction (not a catalyst), acting by a primary impulse.[14] The gel of ferric hydroxide is a good carrier for the Fe^{+++} and Ni^{++} ions in the reaction of oxidation of indigocarmine by means of oxygen of the air at 20°C.[10]

Colloidal carriers have a decisive influence on the structure of the catalyst, conferring upon the adsorbed ions [15] catalytic properties that they generally do not possess in aqueous solutions (except certain hydrolyzed salts).[16] This is contrary to expectation, for dilutions and also the velocity of the ions are greater in aqueous solution than in the adsorbed state on the surface of the carrier. The contradiction can be explained by the fact that during adsorption the ions become rapidly "imprisoned" on the carrier as a result of the formation of an electrostatic field, and their inevitable and strong deformation follows under the influence of individual chemical units of which the gel-carrier is built. There can thus be formed basic salts possessing a structure in a state of formation, otherwise a radical-like structure or a structure of "faint chemical compounds" or *Quasimoleküle*.[17] Artificial charcoal (charcoal from blood, sugar, etc.), with various additions, belongs to this type of catalyst.[18]

REFERENCES

1. cf. Krause, A., *Przemysl Chem.*, 5-6, 267 (1949); 10-11, 558 (1949); see also Haber, F., and Willstätter, R., *Ber.*, 64, 2844 (1931); Waters, W. A., "The Chemistry of Free Radicals, Oxford Univ. Press, Oxford, 1946.
2. Taylor, H. S., *Proc. Roy. Soc.*, London, 108A, 105 (1928); see also Langmuir, I., *J. Am. Chem. Soc.*, 38, 221 (1916).
3. cf. Krause, A., Appelt, K., and Kotkowski, S., *Przemysl Chem.*, 7, 351 (1949); Baudisch, O., and Welo, L., *Chem. Revs.*, 15, 1 (1934).
4. cf. Krause, A., and Krach, H., *Ber.*, 69, 2708 (1936).
5. cf. Andersen-Kraft, C., *Glastech. Ber.*, 9, 577 (1931); Krause, A., and Pilawski, K., *Z. anorg. Chem.*, 197, 306 (1931); and later papers.
6. Fricke, R., and co-workers, *Kolloid-Z.*, 69, 312 (1934); *Z. Elektrochem.*, 40, 630 (1934).
7. Hüttig, G. F., and co-workers, *Z. Elektrochem.*, 41, 527 (1935); Geilmann, W., Klemm, W., and Schüth, W., *Naturwissenschaften*, 20, 639 (1932).
8. Krause, A., and co-workers, *Ber.*, 72, 637 (1939).
9. ———, and ———, *Ber.*, 72, 161 (1939); Schwab, G. M., "Handbuch d. Katalyse," III, Biokatalyse, Wien, 1941.

10. Krause, A., and co-workers, *Farmacja Polska (Warsaw)*, in press.
11. ———, *Przeglad Chem.*, 5-6, 129 (1948).
12. ———, and Hermannówna, A., unpublished.
13. ———, and co-workers, *Problemy (Warsaw)*, in press.
14. cf. Wieland, H., "Uber d. Verlauf d. Oxydationsvorgänge," Stuttgart, 1933.
15. cf. also Kolthoff, I. M., *J. Am. Chem. Soc.*, 58, 449 (1936).
16. Krause, A., and co-workers, *Ber.*, 71, 1033 (1938).
17. cf. Taylor, H. S., and Hüttig, G. F., in Hedvall, I. A., "Reaktionsfähigkeit fester Stoffe," Leipzig, 1938.
18. cf. Warburg, O., *Biochem. Z.*, 119, 134 (1921); Frankenburger, W., "Die Ferment-reaktionen unter dem Gesichtspunkt der heterogenen Katalyse," *Ergeb. Enzym-forsch.*, 3, 1 (1934); Kuhn, R., and Wassermann, A., *Ber.*, 61, 1550 (1928); Nikolajew, L. A., and Kobozew, N. I., *Compt. rend. acad. Sci. U.R.S.S.*, 5, 335 (1947); Kobozew, N. I., and Zuobowicz, I. A., *Compt. rend. acad. sci. U.R.S.S.*, 52, 131 (1946).

PAPER CHROMATOGRAPHY OF AMINO ACIDS

Richard J. Block

Dept. of Biochemistry, New York Medical College, and
Flower and Fifth Avenue Hospitals, New York, N. Y.

and Herbert A. Sober

The National Cancer Institute, National Institutes of Health,
Bethesda, Md.

Introduction

The fundamental processes involved in chromatographic analysis have their counterpart in many natural phenomena. For example, when rain soaks the ground, salts and other diffusable substances tend to become sorted out in the vertical profile of the soil. It has long been known that certain porous substances, such as the diatomaceous earths, charcoal, alumina, etc., are able to take up large amounts of foreign substances both from gases and solutions. The adsorption of gases on charcoal was described by Scheele as early as 1773. In 1785 Lowitz found that charcoal took the coloring matter out of solutions, and at the end of the eighteenth century the adsorption method achieved its first great industrial application in the purification of sugar. One of the first to experiment in isolating delicate substances from complicated mixtures was Danilewski, who in 1862 succeeded in separating amylase from trypsin in pancreatic juice by means of adsorption on freshly precipitated collodion.

Probably the first use of these phenomena for analytical purposes can be credited to Schönbein,[104, 105] who in 1861 proposed the use of starch-potassium iodide paper for the detection of ozone, although lead-acetate paper for detecting hydrogen sulfide had been described in textbooks as early as 1850.[cf29] Schönbein also realized that selective adsorption of the solutes in a solution was taking place at the different heights to which they rose when a strip of filter paper was dipped into the solution. It was Goppelsroeder, however, a pupil of Schönbein, who deliberately took advantage of this selective adsorption and devised a method of *capillary analysis* with which he claimed the detection of 10^{-11} gram of methylene

blue in 1 cc. of solution. Goppelsroeder's first papers were published in 1861 and 1862,[52, 53] and in 1901 a complete monograph appeared.[54] According to this method, one end of strips of adsorptive paper (Schleicher and Schuell No. 598) is placed in a solution of the pigments or other materials to be resolved. As the liquid is drawn into the filter paper by capillary forces, the substances in solution gradually separate from one another, forming a series of bands. The rate of flow of the solvent into the paper strips can be increased by hanging the strips over the edge of a vessel containing the solution of the mixture. To some extent, the bands may be separated from one another by placing the strips in a portion of the fresh solvent after some of the solution has been adsorbed. Goppelsroeder applied his capillary analysis method to the investigation of alkaloids, oils, petroleums, beverages, foods, urine and bile, dyes and to the study of a series of homologous compounds. These studies are ably reviewed by Reinboldt.[95] Because the preparation of materials in quantity is virtually impossible, the method of "capillary analysis" *per se* has found little use in biochemistry, but has been employed extensively by dye chemists following the work of Alexander[2] and others. However, Cochran[30] in 1947 used this method to separate virus proteins from other proteins in plants.

Chromatographic analysis on columns might well be considered the invention of a single individual, Tswett, although he was possibly influenced by the work of Goppelsroeder since he cites this work in his earliest papers. Tswett described the fundamental principles and techniques of chromatography in a remarkable paper in 1906,[126] and his concept of the principles upon which resolution of similar compounds on a column is based, has remained essentially unchanged. In fact, he stated: "It is self evident that the adsorption phenomena described are not restricted to the chlorophyll pigments (with which he made his original observations) and one must assume that all kinds of colored and colorless compounds are subject to the same laws."

As with many great discoveries, the significance of Tswett's observations was not immediately understood and further significant developments and applications for the next 25 years were few. In the United States, Palmer was probably one of the first to carry out extensive systematic studies and to realize the importance of the method.[88] It was not until 1931, when Kuhn, Winterstein and Lederer[65, 66] resolved plant carotene into several components, that chromatographic analysis came into its own. Since that time probably more than a thousand papers have appeared. The multitude of data has been surveyed in recent monographs.[21, 72, 111, 112, 141]

In comparison with the great importance of the chromatographic method, it is rather striking how little we know about its theoretical

background. The chromatographic method is still essentially empirical in character. There are almost no guiding principles of theory to assist in the selection of solvents, adsorbents, or eluents, if the mixture has no apparent points of analogy with mixtures studied by other investigators. Theoretical analysis of the procedures of chromatography,[20, 27, 36, 42, 51, 68, 77, 79, 97, 118, 128, 132] appears to give qualitative rather than quantitative correlation between deduction and experiment. Clarke,[29] discussing selective adsorption and differential diffusion, concludes ". . . developments are retarded by the fact that each is a problem in laboratory manipulation. Some problems cannot always be solved by systematic attack, but must await something akin to inspiration."

Chromatographic analysis is usually carried out by perfusing a mixture of solutes through a column of finely divided adsorbent. If separation takes place, the solutes exist as distinct bands on the column, and in the case of colored compounds, the bands are easily recognized. If fresh solvent is applied to the column, the bands advance at different rates and separation of the bands is increased. At this point, the adsorbent may be extruded from the column, sectioned, and the several bands extracted individually. Alternatively, perfusion of the intact column may be continued with the same or different solvents until each band is washed out of the column and collected separately. The latter technique has been called a liquid chromatogram or elution analysis. New terminology and modified methodology have been introduced by Tiselius.[119, 120] Thus, the method of *frontal analysis* consists of the passage of a solution of solutes *A, B, C, D*, etc., through an adsorbent column until the column becomes saturated and the effluent solution contains the original mixture. The concentration or refractive-index record of the effluent shows a series of increasing steps, the first step consisting of solute *A*, the least adsorbed, the second step consisting of *A+B*, the third step of *A+B+C*, etc. Another method of analysis Tiselius called *displacement development*. In this method, a small amount of the mixture is placed on top of the column, following which a very strongly adsorbed substance is used to push the original mixture off the column. Under these conditions, *A* is followed immediately by *B*, then by *C*, etc., until finally the strongly adsorbed developer appears.

When the components of the mixture are colored, recognition of the separation is relatively simple. However, with amino acids and other colorless substances, special procedures must be used. Thus, the effluent from the chromatogram may be collected in fractions, each of which is then analyzed by conventional methods. Colored derivatives of amino acids, such as the dinitrophenyl amino acids [1, 100] or the copper salts,[15] have proved valuable. Karrer, Keller, and Szony [62] converted amino acids to colored acyl compounds with N-*p*-phenyl-azo-benzoyl chloride, while

the writers have used 3,5-dinitrobenzoyl chloride for the same purpose. Schramm and Primosigh [106, 107] have recommended several methods for making the adsorption of amino acids visible by forming colored compounds on the column by washing with a dilute solution of copper acetate. Gordon, Martin, and Synge [59] have chromatographed acetyl amino acids on columns impregnated with acid-base indicators. The amino acid may also be transformed into a radioactive derivative. [19, 39, 46, 63, 109] Polarographic and enzymatic reactions [61, 75] may also be applied to collected fractions as well as spectrophotometric and bacteriological methods. [135] Successive arbitrary portions of the percolate or effluent may be collected and analyzed gravimetrically, [27] volumetrically, [15, 94, 108] or by specific amino-acid reactions done on a spot-test scale. [10, 40, 43, 82, 84, 108] Claesson [28] has suggested the use of total reflection on a thick glass plate for the detection of invisible zone boundaries. Phillips [90] found that α-amino acids and peptides can be detected on filter paper through their fluorescence by exposure to ultraviolet light; and we have used this technique with great success. In the laboratories of Tiselius, Claesson [27, 121] has developed elegant methods for the continuous refractometric or interferometric analysis of the effluent solution as it emerges from the column. In addition a number of automatic fraction collectors have been described. [16, 94, 108]

Wieland, [129] in 1943, reviewed the chromatographic separations of amino acids under three headings: (1) partition, (2) adsorption, and (3) exchange. Chromatographic separations may be considered in terms of repeated partition of the solute between two immiscible phases, i.e., a counter-current separation. In most forms of counter-current operation, both phases are liquid, and the very small drops required for maximum efficiency cannot be used owing to the difficulty of preventing movement in the wrong direction. However, in true partition chromatography, one of the phases is firmly held on an inert support over which the second immiscible phase is allowed to percolate. This is in effect a liquid-liquid extraction process, a liquid-liquid chromatogram. The second type of chromatographic separation depends on true adsorption in the sense that Tswett used it. In the third type of chromatographic separations employing the zeolites or synthetic ion-exchange resins, resolution is obtained primarily by the charge of the molecule and to a much smaller extent by adsorption, mutual solubility, etc.

The above division of the processes involved in chromatographic separation is somewhat arbitrary and does not hold up when examined carefully but will be maintained at the moment for simplification of further discussion. It has been pointed out by Block [8] that molecular adsorption *per se* plays an important role with synthetic exchange resins, and Mayer, *et al* [79] have considered ion-exchange columns to be analogous to the frac-

tional distillation or extraction column. Kolthoff has pointed out that paper shows ion-exchange reactions with water as a solvent.[112] Freundlich [49] also has stated that adsorption on cellulose is a pure ion-exchange reaction, and concluded that it is not the cellulose itself but rather the impurities like calcium salts that determine adsorption. Recent workers have indicated that with starch [88] or cellulose [26, 122, 123] processes other than partition between liquids are taking place. Finally, Austerweil [4] has suggested that ion-exchange on synthetic ion-exchange resins may be considered a type of partition chromatography.

Developments in Partition Chromatography

Neuberger [87] observed that certain acyl amino acids were extractable from aqueous solutions by chloroform. This observation was used by Synge [113] in 1939, who determined the partition coefficients of a number of acetamino acids between water and water-immiscible solvents such as chloroform, ethyl acetate, etc. As a result of these experiments, it seemed that although the partition coefficients of the acetamino acids were quite similar, separation of one from the other might be accomplished if the distribution between the two phases was repeated often enough in analogy with continuous fractional distillation. Martin and Synge [76] therefore designed an elaborate system of tubes, each tube functioning similarly to a separatory funnel, the ultimate effect of which was the extraction of the acetamino acid from water into chloroform, then from chloroform into fresh water. They constructed a 40-unit counter-current liquid-liquid extraction train which was capable of separating the acetyl derivatives of methionine, valine, proline, the leucines and phenylalanine from each other and from the other amino acids of a protein hydrolyzate. This fractionation train, however, was difficult to construct and to operate, a fact which may have led to the discovery of a new modification of Goppelsroeder's capillary analysis method, called paper partition chromatography.

The counter-current fractionation method between immiscible liquids, however, has been extensively developed by Craig and his co-workers.[6, 36, 60, 131] They succeeded in developing means by which many quantitative extractions can be done rapidly in sequence, and where each step would conform exactly to a term of a binomial expansion. Many substances give a sufficiently constant partition ratio, so that exact mathematical interpretation is possible. Recently,[37] Craig et al. have reported the development of a glass apparatus which permits several hundred separate extractions to be made in sequence in an almost completely automatic apparatus and its preliminary use in the analysis of amino-acid mixtures is indicated.

Martin and Synge,[77] before the development of the Craig apparatus, proceeded on a different line of investigation. They made one phase of the counter-current system immovable by using a mechanical support, such as silica gel, to hold the water, and allowed a second immiscible phase to flow unidirectionally through the water-saturated silica gel. This then is a chromatogram where the separation of components, in theory, depends on the partition between two liquid phases, and not, as in previous chromatograms on differences in adsorption between liquid and solid phases. The potentialities of this type of chromatogram was soon realized and, as is well-known, Catch, Cook and Heilbron [22] shortly thereafter applied silica-gel columns to the separation and purification of the penicillins.

Silica-Gel Method. The silica-gel method has been successfully applied for the separation of a number of the acetylated monoamino acids by Martin and Synge,[77] Gordon et al.,[53-57, 59] Tristram, [124, 125] and others. Unfortunately, the preparation of a uniform and satisfactory silica gel is difficult. The silica gel should function only as a support for the aqueous phase and should not have adsorbent powers of its own. The directions given by Martin and Synge [77] and by Gordon et al.[59] do not seem to be universally applicable, and considerable experimentation often must be undertaken before a suitable preparation is obtained.

In practice, separations achieved are not as good as theory [27, 36, 68, 76, 77, 97] would predict. First, the partition coefficient is rarely constant, usually decreasing as the solution becomes stronger. This results in the front of the band becoming steeper, and the back flatter. At the same time the whole band becomes wider, since the concentrated part moves faster than the dilute portion. On the other hand, an opposing effect may take place at the same time; the more strongly adsorbed solute may replace the less strongly adsorbed one (displacement development), thus tending to sharpen the boundary between the two. In addition, the greater the deviation from a high degree of uniformity of flow throughout the column, the poorer will be the separation.

Paper Chromatography

As a result of the successful separation of certain of the acetylated amino acids on these silica-gel columns, Gordon, Martin, and Synge [58] attempted to separate free amino acids by this means. They found it impractical owing to the adsorption of the amino acids by the silica gel. However, good separations were obtained by using cellulose, in the form of strips of filter paper, as the mechanical support for the aqueous phase. This development has been called paper partition chromatography, par-

tition chromatography, paper chromatogram, papergram, etc. Partition chromatography with paper opened up new vistas for research in many fields, using a technique requiring vanishingly small amounts of material and only simple, cheap, and readily available equipment.

The method of Gordon, Martin, and Synge [55-58] is based upon the principles outlined above for the silica-gel columns, except that filter paper is used as the support for the aqueous phase. In practice, the strip or sheet of filter paper is dipped into a trough containing a water-saturated solution of the developing solvent and the wet solvent is allowed to flow past a spot containing the substances to be chromatographed, which has been applied to the paper previously with a micropipette (i.e., development). Capillarity of the paper pulls the wet solvent from the trough as in the method of capillary analysis (i.e., frontal analysis). The whole operation is carried out in an enclosed space in order to keep the atmosphere saturated with respect to water and the organic solvent.

This method is on the whole very similar to that described by Schönbein [104, 105] and Goppelsroeder,[52-54] except for one important difference. The capillary analysis technique was essentially frontal analysis as the mixture was allowed to run up the paper. The paper chromatograms of Gordon, Martin, and Synge are, in essence, elution analyses or displacement development combined with partition chromatography. It should be recalled, however, that one of the modifications of capillary analysis consisted of preliminary separation of the components in solution on the paper strip, followed by further separation using a pure solvent of a different type.

The paper partition chromatogram method was described in detail by Consden, Gordon and Martin.[31] They reported that, although solvents completely miscible with water can be used, sharper separations are obtained with immiscible solvents saturated with water. However, we, as well as other investigators, have found that completely miscible solutions can be used with equal success. Because of the marked effects of temperature on the solubility of organic solvents in water, the absolute rate of movement of the amino acid with respect to the flow of solvent will vary with the temperature, i.e., the higher the concentration of water, the farther the travel of each amino acid. However, the distance traveled by each amino acid with respect to the others remains remarkably constant over a rather wide temperature change, i.e., although the absolute partition coefficients may be greatly changed, the ratios of the partition coefficients of the respective amino acids are almost unaltered.

R_F Values. The distance traveled by any amino acid on the paper chromatogram is a constant percentage of the distance moved by the advancing front, i.e.,

$$R_F = \frac{\text{distance traveled by amino acid}}{\text{distance traveled by solvent}}$$

Phillips [91] has described a very simple yet elegant method for the rapid measurement of R_F values, which uses an elastic band divided into 100 units. The 100-mark is put on the solvent front and the R_F value for each amino acid is then read off.

The R_F value is a numerical expression which is very constant for each solute under uniform experimental conditions, i.e., type of paper, temperature, solvent system, and to a lesser extent solute concentration. Consden, *et al.*[31] identify the amino acid by its R_F value, but our own experience suggests that greater reliance should be placed on simultaneous controls. Dent [40] has published the R_F map of some sixty amino acids or ninhydrin-reacting substances.

The R_F values for amino acids vary considerably, depending on the type of filter paper used, and in fact vary somewhat from batch to batch of paper. Although the order of magnitude of the values is the same with various papers, there are significant variations in both the absolute and relative amounts which require making new standards with each type of filter paper used. Thus, different papers may be particularly suited for certain special types of procedures.[10]

One- and Two-Dimensional Chromatograms. The relative positions of each amino acid on the finished chromatogram will depend on the solvent system used. Development of the chromatogram in one direction is called a one-dimensional chromatogram. However, if after the development in one direction, with one solvent system, the solvent is removed and the paper is rotated through 90 degrees and another solvent system is allowed to run at right angles to the original direction, than a two-dimensional chromatogram is produced where the amino acids are distributed throughout the sheet, in a specific pattern rather than in only one dimension. Liesegang [69] briefly described this method and called it *cross capillary analysis*. In columns, a comparable separation of solutes can only be achieved by the transfer of various sections of the adsorbent column or fractions of the eluate to a fresh column and rechromatographing with a new solvent system. The two-dimensional chromatogram thus shows a great advantage over any other chromatographic system and the need for such a technique has been ably shown graphically by Dent [40] in Figure 1. The four circles, *A, B, C,* and *D,* represent hypothetical amino acids after a two-dimensional run. One-dimensional chromatograms, run with solvents in the same directions as on the square, are shown below and to the side respectively. In such a case the two-dimensional run would readily resolve a mixture containing some or all of *A, B, C,* and *D.* However, the two one-dimensional runs would show no difference between

mixtures of *A, D; B, C; A, C, B; A, B, D; C, B, D;* and *A, B, C, D*—all would show only two spots on either of the one-dimensional runs.

   Variations in Method. There are several excellent reviews on the methods and scope of paper chromatography,[40, 72-74, 78] to which justice cannot be done here. The experimental modifications of the method of Consden *et al.*[31] are probably legion, since it seems that practically every investigator entering this field has added innovations, although the majority of these seem to have been anticipated by Goppelsroeder.[52-54] The original method of Consden, *et al.*,[31, 32, 39] called descending chromatography, is described as follows: A strip of filter paper containing the amino acids to be chromatographed near the top of the sheet, is allowed to hang

Figure 1. Diagram of paper chromatograms drawn to illustrate the phenomenon of double overlapping. (After Dent.[60])
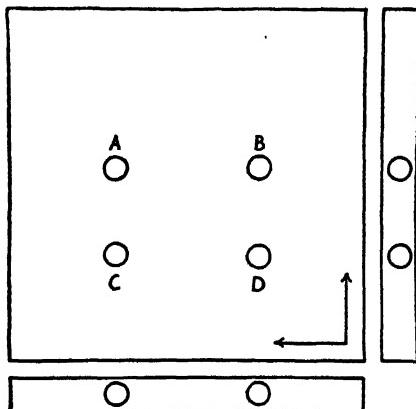
from a trough containing the solvent, and the whole is enclosed in a chamber so that the atmosphere remains saturated with water and the solvent vapor. The solvent is drawn up the paper a short distance by capillarity and then siphons down the suspended strip of paper. Consden[31] used a glass-lined lead box 76×76×13 cm, made tight with a well-fitting lead lid, for his two-dimensional chromatograms. Dent[39] used a large glass and wood box of approximately the same dimensions.

   Williams and Kirby[130] have reintroduced Schöenbein's and Goppelsroeder's method wherein the solvent is allowed to ascend by capillary action. Their apparatus consists of 6-gallon earthenware jars fitted with a glass cover, with the area of contact between jar and cover ground to fit. A dish is filled with solvent and placed in the bottom of the jar. The filter paper (18×11¼ in), after the sample has been added along one edge or one corner, is rolled to form a cylinder, and the edges of the paper are stapled together. The paper cylinder stands unsupported. This apparatus limits the size of two-dimensional chromatograms to somewhat

less than 11×11 inches. The crocks are heavy, porous and retain solvent. Wolfson, et al.[137] have described a somewhat more convenient apparatus for ascending chromatography. Their apparatus consists of a 20-gallon stainless steel tank of 16×24-in inside dimensions, fitted with a heavy glass disc cover and a rubber gasket. The seal is sufficiently tight so that runs may be made under reduced or increased pressure as in the older capillary analysis methods.[95] Twenty-three-inch square sheets may be used in this tank for two-dimensional chromatography, but bending or crumpling of the sheets will occur unless they are rolled into cylinders, stapled, and taped over a double-thickness fold.

For some time[9] we have been using rectangular glass aquaria, which may be purchased at any pet shop. Aquaria, 52 cm long by 32 cm high by 26 cm wide have been used. The lower portion of the sides and all the joints are covered by a layer of paraffin. Two glass rods, running the length of the chamber, are sealed with paraffin to the sides of the chamber about 5 cm out and 2 to 3 cm from the top, and a third rod is sealed half-way between these two and also 2 to 3 cm from the top. Glass troughs are made by sealing the ends of 2-in pyrex tubing and cutting longitudinally down the middle. The chamber is covered by a heavy glass plate and pressed down with a lead weight to make a seal. In operation, three sheets of filter paper 17 to 19-in wide, or smaller, by 19 to 22-in long are cut, the unknown solution applied as drops about 2.5 cm from the bottom of the sheet and 2 to 3 cm apart, and the wet spots air-dried. The troughs are filled with from 20 to 50 ml of developing reagent, depending on the type of filter paper, and the chamber is covered with the heavy glass plate. The solvents climb up the paper by capillary action (ascending chromatography), then past the support rod and down the other side (descending chromatography). The transparency of the chamber permits visualization of the progress of the solvent at all times and without disturbance. The atmosphere within the chamber can be replaced with illuminating gas, nitrogen, etc., as desired, and a beaker containing NH_4OH, HCN, diethylamine, etc., may be placed in the central part of the chamber, if necessary. Excess solvent in the glass troughs serves to keep the atmosphere saturated with respect to the reagents and water.

Three 19×22-inch sheets can be run in each chamber at the same time, i.e., sixty-six one-dimensional runs or three two-dimensional runs in each chamber. For two-dimensional runs, the unknown solution is put on the paper at a point 3 cm from each edge, and then carried through the procedure described above.

The substance to be chromatographed should be put on the paper in as small a volume as convenient. We use 0.01 and 0.005 ml routinely. If large amounts of liquid have to be used, the liquid should be put on the paper with small drops, drying in between drops with a stream

of hot air, or an infrared bulb. Urbach [127] has described an arrangement for the deposition and simultaneous concentration of dilute solutions in paper partition chromatography.

Many other modifications of the original method and apparatus have appeared in the literature.[70, 71, 81, 95, 98, 99, 110, 134] Types of filter paper which are usually employed are Whatman Nos. 1, 3, 4, 54; Schleicher and Schuell Nos. 507, 596, 598, 602, 604.

Solvents. The number of water-saturated solvents that might be used is obviously very great. In our laboratory some 40 to 50 different solvent mixtures were tested and the following were found most generally useful for the separation of amino acids from protein hydrolyzates.

(1) Water-saturated phenol in an atmosphere of coal gas, moistened NaCN and 0.3 per cent NH_3 is used for the separation of alanine, threonine, glycine, serine, cystine, glutamic, aspartic, and cysteic acids in one-dimensional chromatograms. As yet no substitute has been found that will work as well as the first solvent for two-dimensional resolutions. Mallinkrodt's liquefied "Gilt label" product is most satisfactory and is used without purification.

(2) 2,6-Lutidine: ethyl alcohol; water (55:20:25) in the presence of a little diethylamine is another useful reagent. The amino acids in this solvent have a considerably different distribution than with phenol and thus two-dimensional runs with phenol in one direction and the lutidine mixture in the other have proven of greatest value in general exploratory work. Almost every amino acid can be separated on such a chromatogram except for leucine from isoleucine.

(3) Butanol-glacial acetic acid (10:1) in an atmosphere of 0.3 per cent NH_3 again shows a somewhat different distribution in the resolution of the faster-moving amino acids. With this mixture, the spots or bands are particularly sharp, and therefore the method is most useful for quantitative measurement or for cutting out.

(4) sec.-Butanol: tert.-butanol-water (4:1:3) has the same general characteristics as the butanol-acetic acid mixture and, in addition, brings about separation of the slowest-moving components from the intermediate moving amino acids.

(5) Ethanol (77 per cent) shows a good general distribution of the amino acids throughout the chromatogram, is easily removed from the paper, and in particular separates histidine, hydroxyproline and proline very well when used in a two-dimensional chromatogram. Histidine and hydroxyproline sometimes overlap in some other solvent mixtures.

(6) Pyridine (80 per cent) effects a group separation which can then be resolved with other solvents.

(7) n-Butanol: benzyl alcohol (1:1) in an atmosphere of diethylamine effects a separation of amino acids that usually move as a group in other

solvents, i.e., leucine, isoleucine, valine, methionine and phenylalanine, and is used when confirmation of their presence is necessary. A mixture of isobutyric-isovaleric acids (1:1) has been reported by Edman [44] to be especially valuable for the separation of leucine and isoleucine.

By judicious combination of the above solvents in two-dimensional chromatograms and with the help of specific tests for individual amino acids, all the usual amino acids can be identified in a mixture by paper chromatography. Figure 2 shows the expected location of twenty amino acids on a two-dimensional chromatogram as calculated from one-dimen-
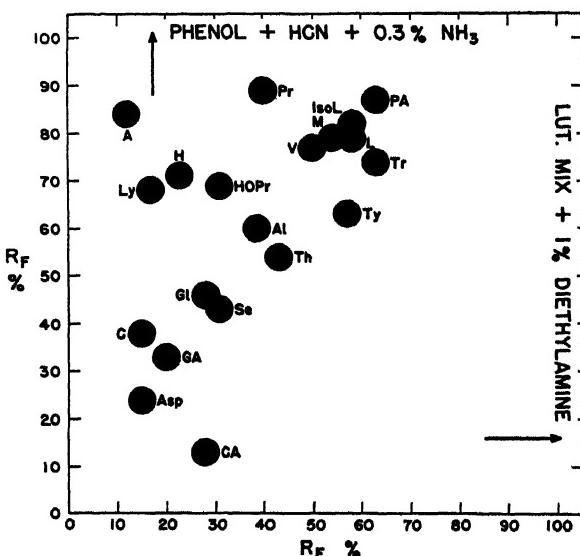
Figure 2. Two-dimensional amino acid distribution.

sional runs. Figure 3, with 80 per cent pyridine instead of the lutidine mixture, indicates somewhat better separation of all the amino acids except for the leucines, valine and methionine. In order to separate the latter group, the butanol-benzyl alcohol mixture in an atmosphere of diethylamine is used (Figure 4). An example of a two-dimensional chromatogram of about 0.3 mg of hydrolyzed casein, run with lutidine mixture, then with phenol, is given in Figure 5.

Spot Identification by Color Reactions. Wherever possible, absolute identification of a spot should not rest solely on the R_F values. In our experience, the position of the spot relative to the other amino acids is most reliable. All the amino acids are revealed by spraying the dried paper with a solution of 0.1 per cent or 0.2 per cent of ninhydrin in butanol, isopropanol, etc. Proline gives a yellow spot, hydroxyproline
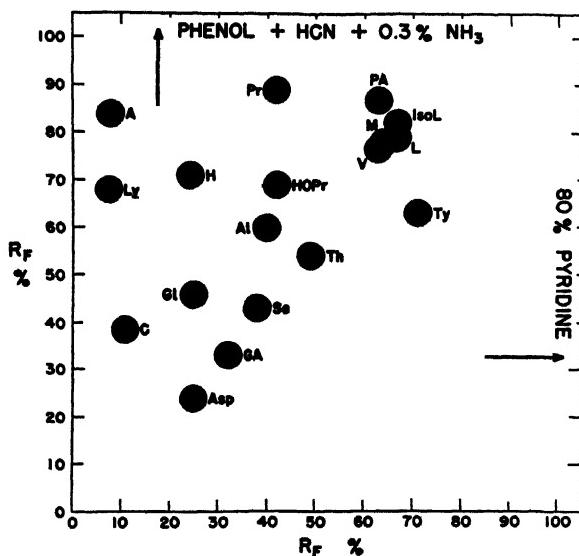
Figure 3. Two-dimensional amino acid distribution.

an orange spot, but all others give blue, purple or reddish-purple spots. Dent [40] reports the detection of as little as 1 microgram of some amino acids on a two-dimensional chromatogram and warns against reliance on R_F values for identification.
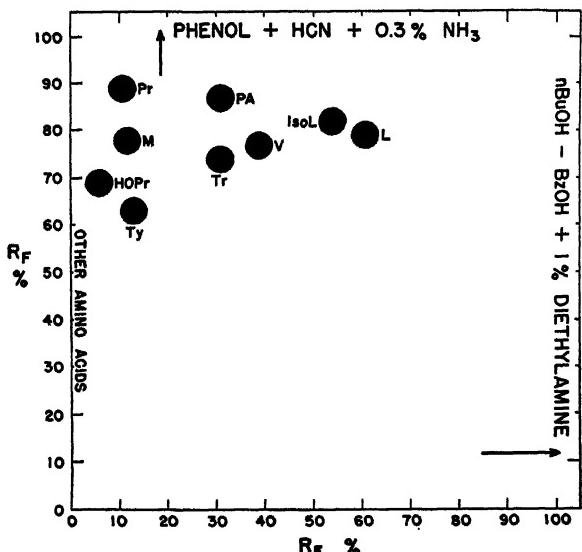
Figure 4. Two-dimensional amino acid distribution.

As far as feasible, the identity of a spot should be confirmed by its movement in a variety of solvent mixtures and by the application of specific tests. During the course of years, a great many color reactions have been developed for certain amino acids.[10, 80] A number of these have given satisfactory results when the developing solvents are removed from the paper. The following amino acids may be identified:

(1) *Arginine* by the Sakaguchi test. The paper is sprayed with 0.1 per cent α-napthol in N NaOH followed by NaOCl ("Clorox").[32, 40]
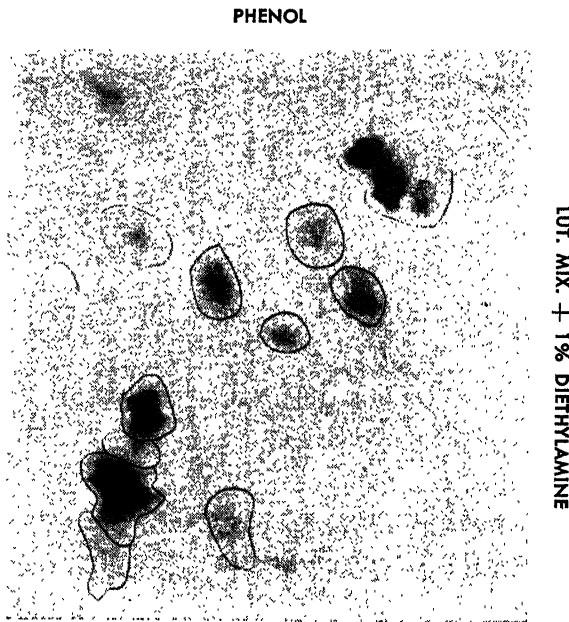
Figure 5. Two-dimensional paper chromatogram of a casein hydrolyzate.

(2) *Histidine* and *tyrosine* by the Pauly reaction. Freshly diazotized sulfanilamide in *n*-butanol is sprayed on the paper and then after 5 minutes drying in air, the color is developed by spraying with saturated Na_2CO_3.[10, 13] No confusion is possible between the two amino acids since they have decidedly different R_f values in almost all solvents.

(3) *Tryptophan*. This amino acid is destroyed during acid hydrolysis and therefore the absence of a spot after acid treatment is indicative of its presence in an alkaline hydrolyzate. It may be detected by spraying the paper with 1 per cent of *p*-dimethylaminobenzaldehyde in 5 per cent HCl.[10]

(4) *Glycine* gives a green color with ortho-phthalaldehyde.[10, 40, 80]

(5) *Proline* and *hydroxyproline* give an intense blue color after spraying with 0.2 per cent isatin in butanol-4 per cent glacial acetic acid on heating.[50] The hydroxyproline color is weaker and greenish and the other amino acids give faint pink colors. The two imino acids are usually well separated from each other on a chromatogram.

(6) *Sulfur amino acids. Cystine, cysteine, methionine,* and some of their oxidized products and peptides may be recognized by their bleaching action when the paper is sprayed with a solution of platinic iodide (equal parts of 0.17 per cent $PtCl_4$ and 0.83 per cent NaI diluted with water (1:10).[32, 133] Sulfur amino acids may also be detected by Feigl iodine-sodium azide reaction.[24] Cystine and cysteine are specifically detected by spraying the dried chromatogram with the Winterstein-Folin uric acid reagent as modified by Block and Bolling.[10]

(7) *Serine* and *threonine*. These hydroxyamino acids liberate ammonia when treated with periodate.[10] Synge has suggested that, when Nessler's reagent is added to solid periodic acid until the precipitate first formed is just dissolved, the resulting solution may be used to identify serine and threonine.[32]

The use of these specific tests aid greatly in confirming the identification of amino acids on paper. X-ray and electron-diffraction techniques on the residue of eluted spots are useful for amino-acid identification.[25]

Quantitative Estimation of Amino Acids

Since the introduction of the paper partition chromatogram, many attempts have been made to estimate quantitatively the separated amino acids. Polson et al.,[92] and Naftalin [86] extracted with acetone the color formed after ninhydrin spraying and determined the amount of substance colorimetrically. The amino acids may also be extracted from the paper and determined colorimetrically [5, 136] or polarographically.[61, 75]

Several methods have been developed which do not require removing the amino acid from the paper. Polson et al.[92] successively diluted known and unknown solutions, and by comparing them, was able to estimate the amino-acid content. Fisher, Parsons, and Morrison [47] found that the area of the spot after ninhydrin spraying was linearly proportional to the log of the concentration. Block,[7] and then Bull [17] reported that if the transmission density through the paper along the chromatogram is plotted against the distance along the strips, then within a limited concentration range, the area under the curve is a function of the concentration (Figures 6 and 7). Fosdick and Blackwell [48] have described a scanning instrument for quantitative one-dimensional chromatography based on the above principle and a continuous recording instrument is now in

construction in our laboratory. Block [7] also found that in two-dimensional chromatography the area of the spot times the maximum color density of the spot also shows a simple relationship to the concentration.

The technically simplest procedure for the estimation of substances on one- and two-dimensional paper chromatograms consists in determining the maximum color density of each spot with an electronic densitometer.[7] The concentration of each amino acid is proportional to its maximum color density and the quantity present on the chromatogram is calculated from an *ad hoc* amino-acid mixture of known concentration; the procedure is carried out simultaneously with the unknown in the same chamber.
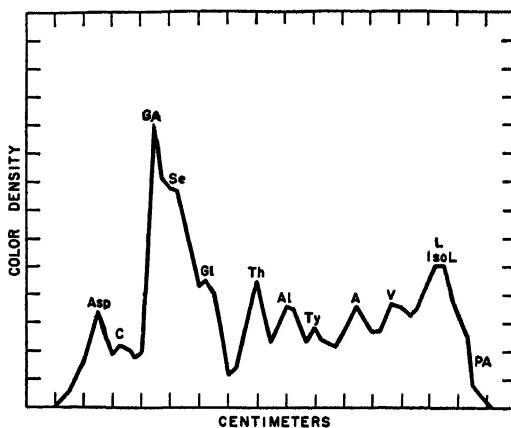
Figure 6. Transmission-density curve; 0.1 mg hair hydrolyzate. Developing solvent was water-saturated phenol plus NaCN and 0.3 per cent NH₃.

Keston, *et al.*[63] have used paper chromatography to separate the *p*-iodophenyl sulfonyl (pipsyl) derivatives of amino acids containing radioactive I[131]. The quantity of each pipsylated amino acid may then be determined from the radioactivity of the isolated areas. Fink, *et al.*,[46] and Stepka, Benson, and Calvin [109] have estimated the number and quantity of radioactively labeled components from radioautographs of paper chromatograms or by direct counting.

Elsden and Synge [45, 114] in 1944 reported that raw potato starch could be used in place of filter paper as the support for the aqueous phase in partition chromatography. Moore and Stein [82-84, 108] used this procedure for the quantitative separation of the amino acids in protein hydrolyzates by dividing the effluent into many small fractions of constant volume by means of an automatic fraction cutter.[108] The fractions are analyzed quantitatively with ninhydrin.[82] The data obtained permit the construc-

tion of effluent concentration curves from which the concentration of each component thus separated is calculated.

Application of Paper Chromatography

Paper chromatography and its modifications have resulted in the reported recognition of new amino acids, namely a, γ-diaminobutyric acid in *Aerosporin;*[23] a-aminobutyric and possibly a-aminoheptylic acids from
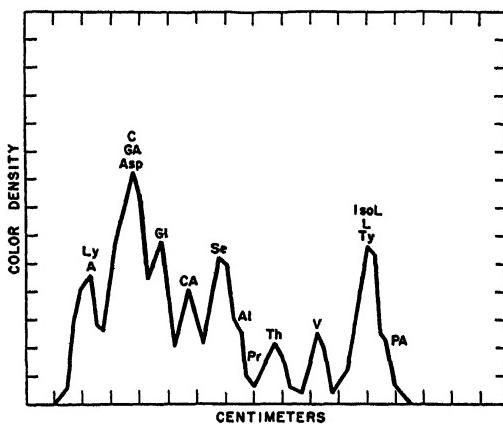
Figure 7. Transmission-density curve; 0.1 mg hair hydrolyzate. Developing solvent was the lutidine mixture plus diethylamine.

hydrolyzate of *E coli,*[92] and a-aminoadipic acid as a degradation product of lysine in guinea-pig homogenates.[14] It has also made possible the investigation of the amino-acid content of proteins and peptides available only in small quantities, i.e., in the salivary glands of *Drosophila,*[67] in bacteriophage,[93] in tobacco mosaic virus,[64] in pollen,[3] in dental enamel and dentin,[12] in pathological urines,[38, 39] etc. It has been used to trace metabolic pathways in photosynthesis;[18, 46, 109] in various phases of growth in the mouse epidermis;[96] in the detection of free amino acids in various tissues;[5, 41] and in the elucidation of structure in salmin,[11, 125] lycomarasin,[139] gramicidin,[35, 101, 115–117] insulin,[100, 102, 140] strepogenin,[138] wool,[32, 33] etc.

For further information on partition and adsorption chromatography, the reader is referred to the reviews by Martin and Synge,[78] Tiselius,[121] Nachod,[85] Strain,[112] Martin,[74] and others.

REFERENCES

1. Abderhalden, E., and Blumberg, P., *Z. physiol. Chem.*, **65**, 318 (1910).
2. Alexander, J., *J. Am. Chem. Soc.*, **39**, 84 (1917).

3. Auclair, J. L., and Jamieson, C. A., *Science,* **108,** 357 (19-18).
4. Austerweil, G., *J. Soc. Chem. Ind.,* **53,** 185 (1934).
5. Awapara, J., *J. Biol. Chem.,* **178,** 113 (1949).
6. Barry, G. T., Gregory, J. D., and Craig, L. C., *J. Biol. Chem.,* **175,** 485 (1948).
7. Block, R. J., *Science,* **108,** 608 (1948).
8. ———, in "Ion Exchange, Theory and Application," ed. F. C. Nachod, p. 296, New York, Academic Press, Inc., 1949.
9. ———, Third Annual Microchemical Symposium, New York, Feb. 27-28, 1948.
10. ———, and Bolling, D., "The Amino Acid Composition of Proteins and Foods," 2nd ed., Charles C. Thomas, Publisher, Springfield, Ill., 1950.
11. ———, Bolling, D., Gershon, H., and Sober, H. A., *Proc. Soc. Exptl. Biol. Med.,* **70,** 494 (1949).
12. ———, Horwitt, M. K., and Bolling, D., *J. Dental Research,* **28,** 518 (1949).
13. Bolling, D., Sober, H. A., and Block, R. J., *Federation Proc.,* **8,** 185 (March, 1949).
14. Borsook, H., Deasy, C. L., Haagen-Smit, C. L., Keighley, A., and Lowy, P. H., *J. Biol. Chem.,* **176,** 1383 (1948).
15. Brazier, M. A. B., *Biochem. J.,* **24,** 1188 (1930).
16. Brimley, R. C., and Snow, A., *J. Sci. Instruments,* **26,** 73 (1949).
17. Bull, H. B., Hahn, J. W., and Baptist, V. H., *J. Am. Chem. Soc.,* **71,** 550 (1949).
18. Calvin, M., and Benson, A. A., *Science,* **109,** 140 (1949).
19. Cannan, R. K., Udenfreund, S., and Keston, A. S., *J. Am. Chem. Soc.,* **68,** 1390 (1946).
20. Cassidy, H. G., *J. Chem. Education,* **23,** 427 (1946).
21. ———, et. al., *Ann. N. Y. Acad. Sci.,* **XLIX** (2), 141-326 (1943).
22. Catch, J. R., Cook, A. H., and Heilbron, I. M., *Nature,* **150,** 633 (1942).
23. ———, and Jones, T. S. G., *Biochem. J.,* **42,** Proc. lii (1948).
24. Chargaff, E., Levine, C., and Green, C., *J. Biol. Chem.,* **175,** 67 (1948).
25. Christ, C. L., Burton, C. J., and Botty, M. C., *Science,* **108,** 91 (1948).
26. Church, H. F., *J. Soc. Chem. Ind.,* **66,** 185 (1934).
27. Claesson, S., *Arkiv. Kemi. Mineral. Geol.,* **23A** (1), 1 (1946).
28. ———, *Nature,* **159,** 708 (1947).
29. Clarke, B. L., in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. **5,** p. 457, New York, Reinhold Publishing Corp., 1944.
30. Cochran, G. W., *Phytopathology,* **37,** 850 (1947).
31. Consden, R., Gordon, A. H., and Martin, A. J. P., *Biochem. J.,* **38,** 224 (1944).
32. ———, ———, and ———, *Biochem. J.,* **40,** 33 (1946).
33. ———, ———, and ———, *Biochem. J.,* **40,** 580 (1946).
34. ———, ———, and ———, *Biochem. J.,* **44,** 548 (1949).
35. ———, ———, ———, and Synge, R. L. M., *Biochem. J.,* **41,** 596 (1947).
36. Craig, L. C., *J. Biol. Chem.,* **155,** 519 (1944).
37. ———, Gregory, J. D., and Barry, G. T., *Cold Spring Harbor Symposia Quant. Biol.,* **XIV** (1950).
38. Dent, C. E., *Lancet,* **2,** 637 (1946).
39. ———, *Biochem. J.,* **41,** 240 (1947).
40. ———, *Biochem. J.,* **43,** 169 (1948).
41. ———, Stepka, E., and Steward, F. C., *Nature,* **160,** 682 (1947).
42. DeVault, D., *J. Am. Chem. Soc.,* **65,** 532 (1943).
43. Drake, B., *Nature,* **160,** 602 (1947).
44. Edman, P., *Arkiv. Kemi. Mineral. Geol.,* **22,** 1 (1945).
45. Elsden, S. R., and Synge, R. L. M., *Biochem. J.,* **38,** Proc. ix (1944).
46. Fink, R. M., and Fink, K., *Science,* **107,** 253 (1948).
47. Fisher, R. B., Parsons, D. S., and Morrison, G. A., *Nature,* **161,** 746 (1948).
48. Fosdick, L. S., and Blackwell, R. Q., *Science,* **109,** 314 (1949).
49. Freundlich, H., "Kapillarchemie," 3rd ed., p. 288, Leipzig, Akademische Verlagsgesellschaft, m.b.H., 1923.
50. Fromageot, C., *Cold Spring Harbor Symposia Quant. Biol.,* **XIV** (1950).
51. Glückauf, E., *J. Chem. Soc.,* 1302 (1947).

52. Goppelsroeder, F., *Verhandl. Naturforsch. Ges. Basel,* **III,** 268 (1861).
53. ———, *Bull. Soc. Ind. de Mulhouse,* **XXXIII** (1862).
54. ———, "Capillaranalyse, beruhend auf Capillaritäts- und Adsorptionerscheinungen", Basel, Berkhäuser, 1901.
55. Gordon, A. H., Martin, A. J. P., and Synge, R. L. M., *Biochem. J.,* **37,** 79 (1943).
56. ———, ———, and ———, *Biochem. J.,* **37,** 92 (1943).
57. ———, ———, and ———, *Biochem. J.,* **37,** 313 (1943).
58. ———, ———, and ———, *Biochem. J.,* **37,** Proc. xiii (1943).
59. ———, ———, and ———, *Biochem. J.,* **38,** 65 (1944).
60. Gregory, J. D., and Craig, L. C., *J. Biol. Chem.,* **172,** 839 (1945).
61. Jones, T. S. G., *Biochem. J.,* **42,** Proc. lix (1948).
62. Karrer, P., Keller, R., and Szony, G., *Helv. Chim. Acta.,* **26,** 38 (1943).
63. Keston, A. S., Udenfreund, S., and Levy, M., *J. Am. Chem. Soc.,* **69,** 3151 (1947).
64. Knight, C. A., *J. Biol. Chem.,* **171,** 297 (1947).
65. Kuhn, R., and Lederer, E., *Ber.,* **64,** 1349 (1931).
66. ———, Winterstein, A., and Lederer, E., *Z. Physiol.,* **197,** 141 (1931).
67. LaCour, L. F., and Drew, R., *Nature,* **159,** 307 (1947).
68. LeRosen, L. A., *J. Am. Chem. Soc.,* **69,** 87 (1947).
69. Liesegang, R. E., *Z. anal. Chem.,* **126,** 172 (1943).
70. Longenecker, W. H., *Science,* **107,** 23 (1948).
71. Ma, R. M., and Fontaine, T. D., *Science,* **110,** 232 (1949).
72. Martin, A. J. P., *Endeavour,* **6,** 21 (1947).
73. ———, *Ann. N. Y. Acad. Sci.,* **49,** 249 (1948).
74. ———, Annual Reports Progress of Chemistry, *London Chem. Soc.,* **45,** 267 (1949).
75. ———, and Mittelman, R., *Biochem. J.,* **43,** 353 (1948).
76. ———, and Synge, R. L. M., *Biochem. J.,* **35,** 91 (1941).
77. ———, and ———, *Biochem. J.,* **35,** 1358 (1941).
78. ———, and ———, "Advances in Protein Chemistry," Vol. II, p. 1, New York, Academic Press, Inc., 1945.
79. Mayer, S. W., and Thompkins, E. R., *J. Am. Chem. Soc.,* **69,** 2866 (1947).
80. Mitchell, H. H., and Hamilton, T. S., "The Biochemistry of the Amino Acids," New York, The Chemical Catalog Co., (Reinhold Publishing Corp.), 1929.
81. Mitchell, H. K., and Haskins, F. A., *Science,* **110,** 278 (1949).
82. Moore, S., and Stein, W. H., *J. Biol. Chem.,* **176,** 367 (1948).
83. ———, and ———, *Ann. N. Y. Acad. Sci.,* **XLIX,** 265 (1948).
84. ———, and ———, *J. Biol. Chem.,* **178,** 53 (1949).
85. Nachod, F. C., "Ion Exchange, Theory and Application," New York, Academic Press, Inc., 1949.
86. Naftalin, L., *Nature,* **161,** 763 (1948).
87. Neuberger, A., *Biochem. J.,* **32,** 1435 (1938).
88. Palmer, L. S., "Carotenoids and Related Pigments," New York, The Chemical Catalog Co. (Reinhold Publishing Corp.), 1922.
89. Patton, A. R., and Foreman, E. M., *Science,* **109,** 339 (1949).
90. Phillips, D. M. P., *Nature,* **161,** 53 (1948).
91. ———, *Nature,* **162,** 29 (1949).
92. Polson, A., Mosley, V. M., and Wyckoff, R. W. G., *Science,* **105,** 603 (1947).
93. ———, and Wyckoff, R. W. G., *Science,* **108,** 501 (1948).
94. Randall, S. S., and Martin, A. J. P., *Biochem. J.,* **44,** Proc. ii (1949).
95. Reinholdt, H., in Houben's "Die Methoden der Organische Chemie," 3rd ed., Vol. 1, p. 291, Leipzig, Georg Thieme, 1925.
96. Roberts, E., and Tishkoff, G. H., *Science,* **109,** 14 (1949).
97. Robinson, F. A., *Analyst,* **71,** 251 (1946).
98. Rockland, L. B., and Dunn, M. S., *Science,* **109,** 539 (1949).
99. Rutter, L., *Nature,* **161,** 435 (1948).
100. Sanger, F., *Biochem. J.,* **39,** 507 (1945).
101. ———, *Biochem. J.,* **40,** 261 (1946).
102. ———, *Biochem. J.,* **44,** 126 (1949).

103. ——, *Cold Spring Harbor Symposia Quant. Biol.*, **XIV** (1950).
104. Schönbein, C. F., *Verhandl. Naturforsch. Ges. Basel,* **III**, 249 (1861).
105. ——, *Verhandl. Naturforsch. Ges. Basel,* **IV**, 1 (1864).
106. Schramm, G., and Primosigh, J., *Ber.,* **77**, 417 (1944).
107. ——, and ——, *Ber.,* **77**, 426 (1944).
108. Stein, W. H., and Moore, S., *J. Biol. Chem.,* **176**, 337 (1948).
109. Stepka, W., Benson, A. A., and Calvin, M., *Science,* **108**, 308 (1948).
110. Steward, F. C., Stepka, W., and Thompson, J. F., *Science,* **107**, 451 (1948).
111. Strain, H. H., "Chromatographic Adsorption Analysis," New York, Interscience
        Publishers, Inc., 1945.
112. Strain, H. H., *Anal. Chem.,* **21**, 75 (1949).
113. Synge, R. L. M., *Biochem. J.,* **33**, 1913 (1939).
114. ——, *Biochem. J.,* **38**, 265 (1944).
115. ——, *Biochem. J.,* **39**, 363 (1945).
116. ——, *Biochem. J.,* **44**, 542 (1949).
117. ——, *Cold Spring Harbor Symposia Quant. Biol.,* **XIV** (1950).
118. Thomas, H. C., *Ann. N. Y. Acad. Sci.,* **XLIX**, 161 (1948).
119. Tiselius, A., *Arkiv. Kemi. Mineral.,* **14B** (22) (1940).
120. ——, *Advances in Colloid Sci.,* **1**, 81 (1942).
121. ——, "Advances in Protein Chemistry," Vol. III, p. 67, New York, Academic
        Press, Inc., 1947.
122. ——, *Arkiv. Kemi. Mineral. Geol.,* **26B**, 1 (1948).
123. ——, *Chem. Eng. News,* **27**, 1041 (1949).
124. Tristram, G. R., *Biochem. J.,* **40**, 721 (1946).
125. ——, *Nature,* **160**, 637 (1947).
126. Tswett, M., *Ber. deut. botan. Ges.,* **24**, 316 (1906).
127. Urbach, K. F., *Science,* **109**, 259 (1949).
128. Weiss, J., *J. Chem. Soc.,* 297 (1943).
129. Wieland, T., *Die Chemie,* **56**, 213 (1943).
130. Williams, R. J., and Kirby, H., *Science,* **107**, 481 (1948).
131. Williamson, B., and Craig, L. C., *J. Biol. Chem.,* **168**, 687 (1947).
132. Wilson, N., *J. Am. Chem. Soc.,* **62**, 1583 (1940).
133. Winegard, H. M., Toennies, G., and Block, R. J., *Science,* **108**, 506 (1948).
134. Winsten, W. A., *Science,* **107**, 605 (1948).
135. ——, and Spark, A. H., *Science,* **106**, 192 (1947).
136. Woiwod, A. J., *Nature,* **161**, 169 (1948).
137. Wolfson, W. Q., Cohn, C., and Devaney, W. A., *Science,* **109**, 541 (1949).
138. Woolley, D. W., *J. Biol. Chem.,* **171**, 443 (1947).
139. ——, *J. Biol. Chem.,* **176**, 1291 (1948).
140. ——, *J. Biol. Chem.,* **179**, 593 (1949).
141. Zechmeister, L., and Cholnoky, L., "Principles and Practices of Chromatography,"
        New York, John Wiley and Sons, Inc., 1941.

In *Chemistry and Industry* of May 20, 1950, S. M. Partridge (Cambridge) discusses displacement chromatography as a preparative method for amino-acids.—*Ed.*

SUCCESSIVE LEVELS OF
MATERIAL STRUCTURE

Jerome Alexander

THE IMPORTANCE of considering all physical and mental phenomena against the background of the material structures involved, necessitates continual reference to the orders of structure existing in nature. In pursuance of this notion, tables showing recognized levels of material structure have appeared in three volumes of this series.[1] Table 1 in the present volume, in addition to outlining the latest available information regarding the various units, from subatomic particles to visible structures, adds two levels not before considered, namely, the *social level,* and the *universal or astronomical level.*

Each of these new levels subsumes all that is contained in its lower levels, though naturally, new orders of facts and relationships emerge. The universal level is included rather for completeness of orientation; but it may be mentioned in passing that the distribution of matter in the universe is not quite as haphazard as might be thought from the random distribution of the stars we see on looking upward on a clear moonless night.

The Astronomical Level

Hubble (Mt. Wilson Observatory of the Carnegie Institution) compares our stellar system to a drifting swarm of bees; our sun with its relatively insignificant family of planets, being a fairly small star in the swarm. The faintest nebulae (i.e., clouds) studied up to now, appear to lie on the surface of a sphere (the visible region) roughly 600-million light years in diameter, a distance which will probably be extended by our newer telescopes. In the present "visible region" are scattered about 100-million nebulae, averaging one-and-one-half-million light years apart, 80-million times brighter than the sun and 800-million times greater in mass. Internebular matter is so small in amount that it does not appre-

ciably dim the most distant nebulae we see, and omitting this from consideration, the concentration of matter in the visible region is only about one gram per 10^{30} cubic centimeters, roughly equivalent to one grain of sand in each volume of space as large as the earth. Such figures adequately express the insignificance of any man, and even of the earth, in terms of matter. The significance of mankind arises from *mentality*, the important emergent factor dominating the *social level* which we will now briefly consider.

The Social Level

The simplest living units show chemical and physical reaction to the milieu, and somewhat higher organisms begin to show tropisms, memory, and apparently purposeful responses. It is hard to say where, in the evolution of mentality, human reasoning power emerged; for mentality exists in and operates through physicochemical systems, so that what we term, respectively, physical and mental phenomena are not separated materially, though they are separated conceptually, that is, in the mind. Many animals lower in the taxonomic scale than apes, monkeys, dogs, and elephants show mental traits which men consider socially desirable, while some men show their animal origin by exhibiting traits regarded as socially undesirable. With the advent of the human mind and reasoning, men slowly acquired the ability of utilizing and directing the operation of natural laws, and of adapting their consequences to human choice, with increasing success. There thus arose an entirely new set of human wants, those of the mind, spirit, soul (the exact word is unimportant), which are involved in what we call human relations (e.g., individuality, affection, love, honor, and respect), and are expressed in philosophy, art, music, literature, etc. These spiritual realities cannot be appraised or measured by any known physical or chemical units,* even though measurements can be made in some of the accompanying physical and chemical mechanisms. To say that these measured changes *are* thought, brings no understanding of the problem which remains as mysterious as ever, despite the unwillingness of some people to admit that there are some things which they cannot explain.

In emerging from the law of the jungle, of tooth and claw, men in their upward and onward advance have striven to establish and enforce rules, regulations and laws, which, in the interest of larger and larger

---

* At the conclusion of a gala concert in Vienna, Emperor Francis Joseph asked his guest, the Shah of Persia, which number he liked best, so that it could be repeated for him. The Shah indicated the first number; but when the orchestra began to play this, the Shah stopped them and said that what he wanted was what they had played *before that*. What had appealed most to him was the tuning up of the orchestra— another illustration of the Latin maxim "De gustibus non est disputandum."

social groups of the gregarious human animal, aim to put limits upon "doing what comes naturally." These laws reflect the *mores*, first of local and then of larger social groups, and are revised from time to time to keep up with advances in human spirituality and idealism. As scientific discoveries made possible increased and low-cost production of food, clothing, and shelter, besides creating and supplying new human wants, the laws respecting property gradually became less rigorous; those protecting the human individual and his rights as against other individuals and as against social groups, including the state, have become, in truly democratic states, more and more extensive and effective. The backward eddies often seen do not indicate the course of the mighty stream of human development.

Allowing full credit to all those who cooperate in the production and distribution of goods, the advance of mankind in supplying its wants, is due mainly to inventors (chemists, engineers, technologists, mathematicians). Despite the great amelioration of working time and conditions, the average man is able to satisfy vastly increased and increasing wants. In the Middle Ages gingerbread was a rarity for kings, and the prayer "Give us this day our daily bread" was for simple subsistence. Today people expect and get much more in return for less labor. Though men differ widely in ability, capability, industriousness and personality, our laws aim to let each person find the niche where he can best serve himself and society, without infringing on the right of others to do likewise. Where free competition exists, the value of a person's service is measured directly or indirectly by what others are willing to pay for the service or its products. Education helps the incompetent, and charity the unfortunate. In operation these principles often suffer because individuals or groups revert to antisocial behavior; but despite this, truly democratic states recognize, proclaim, and increasingly further the principle that all men are entitled to life, liberty and the pursuit of happiness, subject to the like rights of others. Such states do not routinely starve, imprison, or work to death in slave-labor camps those who differ with government officials, elected as servants of all citizens and not as irresponsible and tyrannical masters.

In a letter to Wallace, Darwin said: "The struggles between races of man depend entirely on intellectual and moral qualities." In quoting and commenting upon this remark, Conklin wrote:[2] "Man is not only a biological animal but also an intellectual and social being, and the standards of fitness differ in these three aspects. Biologically the fittest are the most capable of living and leaving offspring; intellectually the fittest are the most rational; socially the fittest are the most ethical. To attempt to measure mental or social fitness by standards of biological fitness is to confuse hopelessly the whole matter and to fail utterly to recognize that

human evolution has progressed in these three directions. Man owes his unique position in nature to this threefold evolution, and while elimination of the unfit is the guiding factor in each of these three lines, the means of elimination and the goals are wholly different."

In discussing "Philosophy, the Guide of Mental Life," [3] the writer stated: "The intimate relation between philosophy and science is evident from the fact that until comparatively recent times scientists were known as natural philosophers. Etymologically, philosophy means the love of wisdom. In its broadest sense, philosophy includes the physical and mathematical sciences, as well as mental and moral philosophy, now often called mental and moral science. Metaphysics (literally, *after physics*) is a branch of philosophy which studies the first principles of being and of knowledge; and though it makes full use of scientific facts, it is often loosely spoken of as 'speculative philosophy'. Students of social phenomena want to be called 'social scientists', for science to-day is a word to conjure with; but they are really social philosophers because they seek not only knowledge but also wisdom in social matters."

Northrop [4] points out that in social science there are two fundamentally different aspects. *Factual* social science aims to determine *what actually is,* whereas *normative* social science aims to establish norms for human and social behavior which may differ from what actually is and which may change from time to time. Law-enforcing agencies are guided mainly by the factual aspects, whereas legislators, judges, and lawyers should be guided mainly by the normative aspects which aim at improving social conditions on the basis of values that are mainly outside the scope of purely physical science. Yet physical conditions are important; e.g., the mental and behavior patterns of identical twins are quite similar. Sociology, involving as it must the exceedingly complex and numerous kinds of relationships between great numbers of persons, races, cultures, and national organizations, all of which are undergoing growth or retrogression, must be guided by philosophic wisdom based on experience but not limited by it. The human mind and society are both undergoing continual evolution.

Religion as a Factor in the Social Level

Practically all racial or political groups having social organization have developed some form of belief as to the origin and the destiny of man. While contact with physical nature led to primitive sciences, e.g., agriculture, manufacture, and the household arts, at the same time love, fear, wonder, and introspection gave rise to primitive philosophies and religions. The early "medicine man" was both scientist and priest, who tried to understand nature and to propitiate the many potent man-like

gods to whom a naive animism attributed the control of rain and crops, lightnings and storms, and the mysteries of life, illness, and death. From these simple beginnings there have emerged a great variety of religious organizations with diverse dogmas and rituals, though they are all based on the general recognition of the limitations of human understanding and faith in the existence of something beyond understanding. Because of the importance of religious beliefs in human social history and progress, we might here consider some of the relations between certain religious ideas and the demonstrated facts of the physical sciences.

Religion and Science. The number of known chemical compounds is very large and continually increasing; but compound formation is always subject to the possibilities inherent in atoms, which are, as we have only recently learned, made up of "primary particles"—electrons, neutrons, protons, etc. Many of our known chemical compounds are not found in nature because the conditions essential to their formation and persistence were not forthcoming. In making heretofore unknown compounds, chemists *create* nothing; they simply find conditions under which certain atomic possibilities can establish themselves. The same is the case in nuclear transformations, where the possibilities inherent in subnuclear units establish themselves. Scientists merely change the relative positions of various material units or mechanisms, and then observe what happens according to natural laws. *Actual creation* is utterly inscrutable, and beyond human power.

The Creation, however, established in matter certain inalienable potentialities. Under conditions existing on our relatively tiny earth, the "fundamental" subnuclear units are capable of forming only about 650 more-or-less stable atomic nuclei which attract satellite electrons to make up what chemists have long termed the ninety-odd "chemical elements." In turn, as every chemist well knows, these "elements" can combine only in certain proportions to make more-or-less definite and stable molecules, the next higher order in the hierarchy of material aggregations. Thus in the case of hydrogen and oxygen we know only the ubiquitous and stable water (H_2O), and the relatively scarce and unstable peroxide of hydrogen (H_2O_2); the mass spectrometer demonstrates the transient existence of OH. Even the most complicated combinations of atoms and molecules found in nature or assembled in the laboratory exist only because of properties inherent in matter since its creation. Matter follows its own laws, which are not necessarily those of calculators.

To explain a phenomenon, we must descend to the level of material structure where comprehensible basic activities emerge, so as to make clear what J. Clerk Maxwell termed the "go" of it. However, the units used in our explanation may themselves be complex and subject to variation. But names are often given to phenomena long before they can

206

TABLE 1. SUCCESSIVE STRUCTURAL LEVELS

Material Structures	Order of Complexity	Mode of Examination	Approximate Size	Chemical Units	Biological Units	Cotton	Starch	Carbon Steel
Atomic nuclei....	1	Transmutation, cyclotron	He = 8×10^{-5}Å, Au = 1×10^{-4}Å	C, H, O, N, P, S, Fe, K, Na, Ca, Mg, Cl, Fl, B, I, Zn, Cu, Cr, Ni, V		C, H, O	C, H, O	Fe, C(0.1 – 1.0%) (S, P, Si, Mn, etc.) 0.5% C = ca 7.5% FeC_4
Atoms..........	2	Cyclotron, x-ray, chemical	5 Å±					
Molecules........	3	Infrared and Raman spectra, x-ray, chemical	50 Å±	Amino acids, lipoids, nucleotides, glucose, etc.	Vitamins, hormones, pyrophosphate, carriers, prosthetic groups	Glucose, etc.	Glucose	Fe, C, FeC_4
Macromolecules*. Molecular aggregates.....	4	Electron microscope, ultramicroscope, chemical	100 mμ±	Proteins, glycogen, cellulose, starch, etc.	Enzymes,† chlorophyll, cytochrome, bacteriophages, genes, viruses, etc.	Molecular chains and groups	Molecular chains and groups	α-iron (Ferrite), γ-iron, Cementite
Micells..........	5	Electron microscope, ultramicroscope	1 μ±		Cytoplasmic structures	Aggregates with impurities, Fibrils		
Microscopically resolvable units	6	Chemical, microscope	¼ μ±		Chromosomes, nuclei cells, bacteria		Starch grains	
Visually resolvable units.........	7	Eye	50 μ±		Tissues, organs, drosophila, mouse, whale	Fibers, yarns, fabrics, etc.		Austenite (α-iron), Martensite (γ-iron), troostite, sorbite, pearlite, Hair-spring, saw, engine, bridge, sky-scraper
Social level......	8				Man			
Universal level....	9	Telescope	2 billion light years in diam.					

* Primary colloidal particles, 30 mμ±; Secondary colloidal particles, 100 mμ±. † Moleculobiont, hypothetical simplest living unit.

TABLE 1. SUCCESSIVE STRUCTURAL LEVELS

Fundamental Particles

Order of Complexity 0 (?)	Charge	Size* (Å)	Mass (gram)	Discoverer	Where Found
Electron.............	$-e$	3×10^{-5}	9×10^{-28}	Sir J. J. Thomson (G. Johnstone Stoney) (Sir Wm. Crookes)	Electric current Negative electrostatic charges Nuclear satellites and emissions
Positron.............	$+e$	3×10^{-5}	9×10^{-28}	C. D. Anderson	Emitted by atomic nuclei
Proton..............	$+e$	2×10^{-8}	1.66×10^{-24}	Sir Ernest Rutherford	Nucleus of protium (H)
Neutron.............	0	2×10^{-8}		Sir James Chadwick	Atomic nuclei

Mesons, formerly called "mesotrons," were discovered in cosmic radiation by C. D. Anderson and S. Neddermeyer (1936). Though commonly listed as "fundamental particles," they seem to be short-lived entities. There is some evidence that a neutral meson exists, but the two kinds now recognized have either positive or negative charge, and differ both in mass and persistence. The π-meson has a mass 286 times that of the electron and a life of about 10^{-8} second. The μ-meson has a mass 215 times that of the electron, and a life of about 2.2×10^{-6} second (*data from C. F. Powell, Bristol, England*).

Quanta of Interaction now recognized as providing distinct means for the transfer of energy between one elementary particle and another:

Photon: Elementary unit of radiation energy carried away by the "electromagnetic field" from a radiating atom or other system of particles.

Neutrino: Elementary unit of energy given off by a radioactive nucleus in every decay process in which simultaneously an electron or a positron is released.

* The uncertainty principle makes all these measurements indefinite.

be explained in terms of material units. Such names are commonly regarded as "metaphysical" until an acceptable explanation is found—whereupon they may be gradually accepted as "scientific." The natural mental inertia to the acceptance of new ideas is intensified when there is also involved the abandonment of long-taught but erroneous notions, such as the indivisibility of atoms, or the belief that life can be accounted for only by assuming a series of special creations separate and distinct from the incomprehensible initial Creation. Hippocrates called the unknown directive force of life "physis" (nature). Aristotle, and later Driesch, called it "entelechy." Paracelsus called it "archaeus." Bergson called it "élan vital." However comforting these words may have been to philosophers, they suggested no *material* basis for understanding the very real phenomena to which they referred, and they were therefore anathema to men of science.

Scientists have been endeavoring to understand how life originated and slowly developed into mankind by the working of natural laws. Viewed from this standpoint, science has real religious significance, entirely apart from any dogma or theology; for by studying the ways in which matter has worked out or can work out its possibilities in animate and inanimate nature, scientists are revealing many unsuspected and wonderful consequences of the Creation. The billions of years which it took to evolve the earth, and eventually, man, do not bulk large in eternity; but the evolutionary process gives a reasonable explanation of the statement in Genesis, II, 7: "The Lord God formed man of the dust of the ground and breathed into his nostrils the breath of life. And man became a living soul." If literally interpreted, this poetical expression could be taken to mean that God made a mud-pie and blew on it. Such an anthropomorphic interpretation is inconsistent not only with known facts, but also with a mature and exalted concept of Divinity. It is unfair to God; whereas science continues to reveal the wonderful and intricate mechanisms set into operation by the Creation. Perhaps the most wonderful result is that with the evolution of the human body there came the equally real evolutionary development of sensitivity, mind, and soul.

REFERENCES

1. Alexander, J., in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vols. I, II, V, New York, Reinhold Publishing Corp., 1926, 1928, 1944.
2. Conklin, E. G., *Bios,* 19, 254 (1948).
3. Alexander, J., "Life—Its Nature and Origin," Chap. 12, New York, Reinhold Publishing Corp., 1948.
4. Northrop, F. S. C., "The Meeting of East and West," p. 256, New York, Macmillan Co., 1947.

Part II

BIOLOGY AND MEDICINE

SURFACE CHEMISTRY AND BIOLOGY

A. E. Alexander

Dept. of Colloid Science, Cambridge University, Cambridge, England

Introduction

The part played by surface phenomena in biological systems has been pointed out too often to need emphasis here. Realization of its importance has, in the past, led a number of biologists to study surface chemistry, as Devaux and Hardy, both physiologists, who nevertheless made very substantial contributions to physicochemical phenomena in this field.

The differentiation of function in even the simplest biological system, such as a bacterium or red cell, necessitates the existence of regions of markedly different composition, one of the simplest cases being a lipoid phase in contact with an aqueous phase containing soluble proteins. The movement of chemical substances, whether nutrients or drugs, will thus as a rule involve adsorption, or uptake in some form, by the bounding surfaces or interfaces. Many compounds normally present in cells, and also many drugs, have pronounced surface-active properties, and by their tendency to be adsorbed may modify the cell surface as well as the penetration through it of other molecules.

In view of the complexity of even the simplest biological system—an isolated cell for example—little hope of any detailed understanding of the physicochemical factors at work would be likely in the absence of data upon much simpler systems. The field of *surface chemistry* covers the physicochemical study of adsorbed and insoluble films at all the possible types of interface, particularly the air-water, the oil-water, the solid-liquid, and the gas-solid. Most of this work has some relevance for the biologist, although that from the first two types, the air-water and oil-water interfaces, is naturally of more immediate application. The present review aims at indicating, chiefly by means of selected examples, the principal ways in which surface chemistry has been of value.

The basic techniques used in surface work involve measuring the change in surface tension (termed the *surface pressure*), the change in the phase-boundary potential (termed the *surface potential*) and the change in viscosity of the surface layer (the *surface viscosity*); in certain cases electrophoresis of discrete particles forms a useful adjunct. Details of such measurements, which require comparatively simple apparatus, can be found in suitable books.[1, 2]

Determination of the Structure of Complex Organic Molecules

Of historical interest, and marking a major advance in the development of surface chemistry, was the study of insoluble monolayers of natural oils (e.g., olive oil), by Pockels and Rayleigh at the end of the last century, the latter showing how the molecular size could be com-
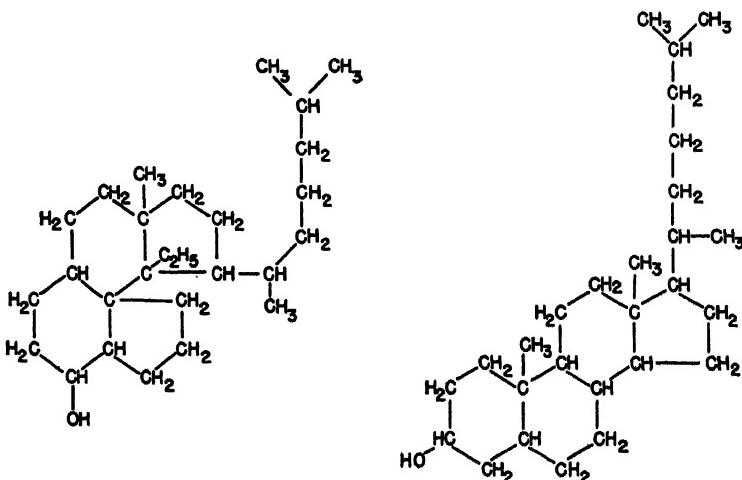
Figure 1. Structures of cholestanol, the old formulation on the left, the modern on the right.

puted from the amount of oil necessary to give a measurable surface pressure. More recently, the value of surface films was first emphasized by the work of Adam and his co-workers with the sterols.[1]

In general terms it can be said that the technique can indicate not only the general shape of the molecule, but also the relative position of the polar groups if more than one happens to be present. (One polar group at least is essential to ensure adequate spreading.) Thus, a molecule consisting of a complex ring system with two polar groups in close proximity would tend to give a condensed film, whereas if these groups are widely separated a gaseous or expanded film is more likely,

owing to the resulting tendency of the molecule to lie flat on the surface. Where condensed films are formed, it is possible to compare the area at zero compression with that calculated from models of likely structures.

The work of Adam on the sterols, mentioned above, showed that the basic skeleton was much more elongated than was believed at the time. The limiting film area of 38 to 40Å^2/molecule given by many members (e.g., cholesterol) agrees reasonably with that calculated from the modern structure; the older formulation would have given one of 50 to 55Å^2. The difference between the two is clearly seen from Figure 1, the molecule being anchored to the water surface by the —OH group. Other compounds whose constitutions were studied about the same time were batyl and chimyl alcohols,[3] and certain porphyrins and related compounds, such as hemin, protoporphyrin, chlorophylls and phthalocyanines.[4] In

Figure 2. The structure of hedraganic acid and the position of the carboxyl group. Position 17, Haworth and Ruzicka; position 20, Kon and Ross.
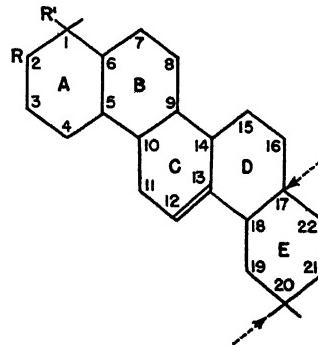
the case of natural batyl alcohol, for example, the force-area curves coincided with those from synthetic α-octadecyl glyceryl ether, giving strong support for the identity of the two compounds.

Recent work along similar lines has dealt with such problems as the structure of cerin, friedelin and related compounds,[5] lupane derivatives,[6] the constitution of quillaic and oleanolic acids,[7] and the position of the carboxyl group in certain triterpene acids (sapogenins).[8] Figure 2 shows the basic skeleton of this last type of compound, where $R'=R=H$ in hedraganic acid, and $R'=Me$, $R=H$ in the oleananic acids. The force-area curve given by hedraganic acid shows a limiting area in the condensed state of just over 40Å^2, a value much more consistent with the —COOH group being in position 20 than in the 17 position as formulated previously.

Stenhagen and his co-workers[9] have, over the past decade, made detailed monolayer studies of a number of branched-chain compounds, such as phthioic acid and phthiocerol, and by comparison with synthetic

compounds have thrown a good deal of light upon these biologically-important compounds.

Proteins, although frequently very soluble in water, can nevertheless be spread quantitatively at both air-water and oil-water interfaces. As would be expected, these surface layers have evoked much interest, both as a means of elucidating molecular structure and from the viewpoint of the formation and structure of protein membranes.

Spreading is known to be accompanied by a complete loss of the original water solubility (*surface denaturation*), the constituent polypeptide chains being unfolded and reorientated at the surface. The monolayers combine high compressibility with marked elastic and hysteresis effects, but any detailed analysis has only recently become possible following a study of monolayers of polymers of known structure.[10] The principal synthetic polymers examined have been:

polyacrylates, $\left(\begin{array}{c} -CH_2-CH- \\ | \\ COOR \end{array}\right)_n$, where R = Me, Et, Bu, etc.,

polymethacrylates, $\left(\begin{array}{c} CH_3 \\ | \\ -CH_2-C- \\ | \\ COOR \end{array}\right)_n$, where R = Me, Et, Bu, etc.,

nylons, eg. $(-(CH_2)_6 \cdot CO \cdot NH-)_n$, and quite recently the

polypeptides,[11] $(-CO \cdot NH \cdot CHR-)_n$, where R = $-CH_3$, $-CH_2 \cdot C_6H_5$, etc.

Many of these show similarities to protein monolayers, so that the characteristic properties of the latter certainly arise in part from the entanglements between the long polypeptide chains. Figure 3 presents some surface data for a typical protein and for a synthetic polypeptide at the oil-water interface.[12] The steep rise in the surface viscosity probably indicates the point of close-packing.

Another factor which seems to be important in protein monolayers is hydrogen bonding between the $=C=O$, $=N-H$ groups in adjacent chains, for the study of long-chain compounds containing this group (e.g., amides, ureas), shows strong intermolecular linkages.[13] Such intermolecular hydrogen bonding is usually broken down by very strong acids (e.g., $4N \cdot H_2SO_4$), the original solid film liquefying. In the case of the synthetic polypeptides,[11] probably owing to their regular structure, the bonding is sufficiently strong to resist even such drastic measures.

By means of monolayers of proteins, synthetic polypeptides and other polymers, long-chain amides, etc., it is possible to examine in some detail the effects of agents known to influence native proteins, such as urea,

thiocyanates, etc. Work along these lines is in progress in the laboratory
at Cambridge.

One further point concerning protein monolayers deserves mention—
namely the possibility of measuring molecular weights by this means.[14]
The basic principle is very simple—merely measure the force-area curve
in the dilute gaseous region and extrapolate to zero pressure, where the
ideal gas equation $\pi A = kT$ should be obeyed ($\pi =$ film pressure, $A =$ area).
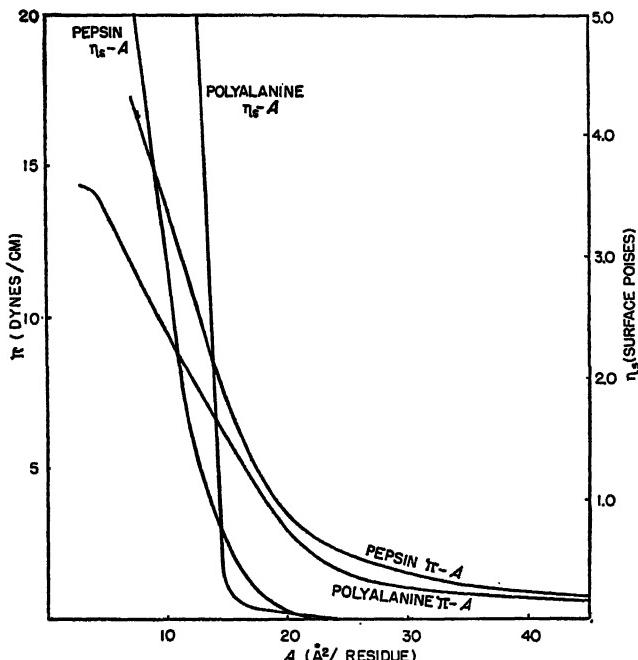
Figure 3. Force-area and viscosity curves for the synthetic polypeptide
polyalanine, and for a typical protein (pepsin) at an oil-water inter-
face.

Experimentally it is not at all easy, owing to the minute pressures in-
volved. Guastalla's values for ovalbumin and gliadin agreed reasonably
with those from ultracentrifuge data, whereas the results for hemoglobin
indicated splitting into five or six fragments. Allan [15] has recently ex-
amined the problem in detail, using monolayers at both air-water and
oil-water interfaces, and synthetic polymers as well as various proteins.
With proteins spreading from aqueous solution upon strong salt solu-
tions (e.g., 25 per cent $(NH_4)_2 SO_4$) gave the best results. The molecular
weight thus found for egg albumin is of the same order as the bulk value,
whereas that for insulin indicates splitting into four or five fragments.

Reactions at Interfaces

Of the great variety of reactions at interfaces which have been studied, we shall restrict ourselves to those of rather obvious biological interest. Such reactions range from *irreversible* chemical processes involving the rupture of chemical bonds, to interactions which are readily reversible, and which are therefore regarded as being more physical in nature.

Some of the simpler physical systems involve studying the effects of water-soluble substances (drugs, dyes, etc.) upon monolayers of proteins, cholesterol, phospholipoids, etc. Interaction is measured quantitatively by changes in surface pressure, surface potential, viscosity, etc., and the strength of attachment to particular cell components can thus be detected. Such studies are clearly relevant to an understanding of the interplay of biologically-active compounds with cell membranes, which although often only the first step in a chain of processes, may nevertheless be of determining importance. The agglutination and lysis of red cells, the outer membrane of which is thought to consist largely of proteins and lipoids (e.g., cholesterol), has thus been examined.[16] With a series of water-soluble soaps $C_{12}H_{25}X$ ($X = $ —COONa, —SO$_4$Na, etc.), the relative degree of penetration into monolayers of cholesterol or protein was found to parallel their hemolytic efficiency. On the other hand, multipolar compounds such as tannic acid, silicic acid, and certain dyes such as Janus Green, which agglutinate rather than hemolyze, show no penetration but are strongly adsorbed beneath protein films. These results have been regarded as confirmatory evidence for the lipo-protein theory of the red-cell membrane.

One of the simplest interfacial reactions—the adsorption of proteins—has evoked much interest, particularly in connection with enzymes and the antibody-antigen reaction. Spreading at an *air-water* interface seems to be accompanied by the complete loss of enzymatic and specific immunological functions,[17, 18] as well as of solubility, indicating very radical and irreversible changes in structure. In the technique usually employed the protein, after being spread as a monolayer, is removed to a metal plate by a dipping process, and then tested for enzymatic activity. Where some activity has been reported, as with catalase and saccharase, it seems significant that the film thicknesses are all very much greater than that for a fully spread monolayer (ca. 10A), indicating the presence of some unspread protein.[18]

At *oil-water* interfaces, the extent and the rate of surface denaturation appear to be markedly less than at the air-water interface.[12, 19] The adsorption diminishes as the interfacial tension is reduced; electrical effects arising from the charges on the interface and the protein are also important.[20] Considering the importance of these adsorbed protein films as

stabilizers of biological emulsions, as precursors of biological membranes, and in immunological reactions, comparatively little is known about their detailed structure.

As previously mentioned, the spreading of an antibody or an antigen (if protein in nature) at an air-water interface appears to be accompanied by the loss of the specific immunological reaction. For example, horse-serum globulin as monolayer and rabbit antiserum in the subsolution, and pneumococcus (Type II) antibody as monolayer and specific polysaccharide in the subsolution, gave no detectable reaction with surface-pressure and surface-potential techniques.[21] It would be of great interest to carry out similar investigations at oil-water interfaces of various types.

It has been known for some time that multilayers (built-up films), such as barium stearate, after *conditioning* by immersion into aluminium or thorium salts (to deposit a thin layer of hydroxide) will adsorb proteins from solution. The thickness of the adsorbed film can readily be measured from the change in the optical properties (color) of the film. A film of an antigen deposited in this way will specifically adsorb its homologous antibody from solution, and vice versa. In such systems it seems clear, from the increments in thickness, that the adsorbed proteins are in the globular state, so that the preservation of specific properties is not surprising.

Recently, the effect of various coatings upon antigen or antibody films deposited on a *conditioned* plate in the above way has been examined.[17] The film of an antigen (for instance) is coated with a multilayer of barium stearate or by a layer of Formvar (a synthetic polymer). The plate is then immersed in a solution of the homologous antibody, and the ensuing adsorption is measured optically from the increase in thickness. By this means appreciable uptake has been observed even though the deposited film is many molecules thick, and this has been held to indicate the existence of long-range forces in specific immunological reactions.[17] The argument assumes that the layers on the plate, both protein and the polymer or multilayer, possess perfect laminated structures. From the known structure of multilayers and from the behavior of proteins at interfaces, such a high degree of perfection would seem most improbable; and as evidence for the existence of long-range specific forces in immunological reactions, the observations are by no means convincing.

An illustration of the power of surface methods is provided by the study of the photochemical breakdown of proteins.[22] Proteins undergo photochemical hydrolysis when irradiated by ultraviolet light; the reaction can be followed very conveniently in a protein monolayer illuminated from above, thus overcoming troubles caused by absorption of radiation by the solvent. Preliminary work using paraffin-chain amides showed the importance of an aromatic nucleus as chromophoric group:

$$C_{17}H_{35} \cdot CO \cdot NH \langle \bigcirc \rangle + H_2O + h\nu \rightarrow C_{17}H_{35}COOH + H_2N \langle \bigcirc \rangle$$
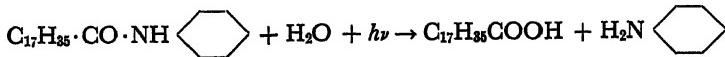
With monolayers of proteins the extent of breakdown seemed to be determined by the number of aromatic side-chains, in conformity with the above conclusion.

Biological Activity in Homologous Series

Theories of drug action, particularly those concerned with the relative potency of members in an homologous series, have been the subject of much bitter controversy in the past. Overton and Meyer, and their school, in their explanation of narcosis, laid supreme importance upon the solution of drugs in cell lipoids, whereas Traube considered that capillary activity, i.e., adsorption on cell surfaces, is paramount.

It is frequently found that the biological response to a homologous series of drugs bears an exponential relationship to the number of carbon atoms in the hydrocarbon chain. Many physical properties such as surface activity (Traube's rule) and fat solubility show just the same exponential relationship, so that it is usually impossible to infer the mode of action of the drug from any simple correlation between biological action and a particular physical property (see Ferguson [23]). All that can be said is that the slowest physical process in the over-all effect depends on a partition property of the molecule—the particular one involved cannot readily be specified, and various partition effects may be involved in the complete process of drug action.

Trim and the author [24] have recently reconsidered the problem, particularly in relation to the influence of soaps upon drugs (as outlined below), which throws some light on the problem. They concluded that compounds such as hydrocarbons, both simple and chlorinated, are relatively devoid of surface activity and do not act by a Traube mechanism. On the other hand, with many drug series, such as the n-alkyl phenols, alkyl resorcinols, etc., the amount of adsorption at the cell surfaces (e.g., with bacteria), seems to be the decisive factor in the overall biological activity.

Numerous attempts to correlate surface activity with biological activity have appeared in the literature. The former has generally been assessed from the lowering of surface tension (i.e., at the air-water interface), less frequently from interfacial tension against an oil such as medicinal paraffin. Of these the latter is undoubtedly the better, although it must be emphasized that the only completely satisfactory system is the biological surface itself which, unfortunately, is usually a matter of some difficulty. The chief difference between adsorption upon a biological and an

air-water or oil-water interface arises with ionized drugs and when the biological surface and drug ion are oppositely charged.

It should also be pointed out that most biological activities refer to dynamic rather than equilibrium conditions. Under such conditions the processes of adsorption and *ab*sorption (i.e., solution in the lipoid layers of the previously *ad*sorbed molecules) appear to be the rate-controlling steps with most surface-active drugs.[24] The importance of surface activity is thus apparent.

Many apparent exceptions to Traube's rule have been reported, even in series such as the *n*-alkyl phenols, resorcinols, etc., where surface activity is undoubtedly of major importance. Most of these anomalies arise from solubility effects,[24] others because the basic assumptions of Traube's rule are not valid.[25] Of these assumptions, the most important is that the amount of adsorption is very small, which may not be the case in some systems.

Biological Activity of Soaps (Natural and Synthetic)

Soaps, which together with the dyes constitute the important group of colloidal electrolytes, may be taken as the extreme case of ionized organic drugs, combining a lipophilic hydrocarbon chain with a strongly hydrophilic polar group. They are also an extreme type of drug in the sense that their lipoid solubility is in general very low, and almost all their biological activity appears to arise, directly or indirectly, from their pronounced adsorptive powers. Colloidal electrolytes include the natural and synthetic soaps (detergents and wetting agents), many dyes, most bile salts, certain phospholipoids such as lecithin, and a few ionized drugs containing long paraffin chains. Both the surface activity and the tendency to aggregate (micelle formation), which to-gether are responsible for most of the biological effects of these substances, arise from the tendency of the paraffin chains to be expelled from the aqueous phase. Micelle formation sets a very effective limit to adsorption, as thereafter the boundary tension is but little affected.

The Traube theory of drug action seems to provide an adequate explanation in almost all cases.[24] As the surface activity, measured by the lowering of surface or interfacial tension, increases, so in many cases and in a general way does the biological activity. The apparent discrepancies all seem to involve conditions where the organic ion of the drug and the biological surface are oppositely charged: in such cases electrical (Coulomb) forces will increase the adsorption above that anticipated from measurments at an air-water or oil-water interface.

In addition to concentration and charge effects, adsorption is also influenced by the structure of the molecule, as has been shown in studies

of fatty acids and synthetic soaps. With these ionized compounds the branched-chain isomers are less surface active than the straight-chain compound at low concentration, but there is frequently a reversal at high concentrations. The difference arises from the greater difficulty experienced by the branched-chain isomers in packing together to form micelles,[26] which, as mentioned above, sets a very effective limit to surface activity. Another contributory factor arises from the dissociation constant not being identical in the various isomers:[27] this may modify the adsorption since most comparisons are done at a fixed pH.

Salts tend to increase the adsorption of soaps since they reduce the Coulombic repulsion between the adsorbed organic ions and those in solution. As would be expected the effect is caused primarily by the concentration and valence of the ion oppositely charged to the organic ion— thus, in the case of anionic soaps, $CaCl_2$ is more effective than $NaCl$.[28] The addition of larger amounts of salts usually brings about precipitation of soaps and the effects upon the surface tension are again reflected in the biological activity. Many examples of synergistic action in mixed drug systems are also explicable in terms of changes in surface activity.[24]

The physical and chemical nature of the adsorbing surface also influences the adsorption of soaps, and this has been used to explore the nature of the bacterial envelope.[29] Many effects of pH, salts, and of bacterial type (Gram-positive or Gram-negative) upon the bacteriological activity of anionic and cationic soaps can be explained by paying due regard to the charges on the biological surface.

Soaps, as previously mentioned, aggregate to micelles above a certain rather definite concentration, the *critical concentration for micelles.* Owing to their peculiar structure, micelles interact very strongly with, and thus indirectly influence the biological activity of, a wide variety of substances, ranging from water-insoluble hydrocarbons at the one extreme, through partially soluble organic molecules like phenol, to the very hydrophilic, simple inorganic ions. The way in which micelles adsorb and thus detoxicate phenols will be discussed in the next section.

The uptake of sparingly-soluble substance by micelles is usually termed *solubilization.* It provides a mechanism for transporting sparingly-soluble substances, both of drugs and of lipoid constituents from biological surfaces. The latter effect would explain the sudden increase in permeability observed in some systems above the critical concentration for micelles.[24]

Finally, mention must be made of the interaction of soaps with substances frequently present in biological tests, particularly polymers such as proteins, peptone, agar, etc. Measurement of the surface activity of such soap-polymer systems indicates that the polymer tends to adsorb the soap, the effect being particularly marked if oppositely charged, e.g., cationic soaps with gelatin or agar at neutral pH. This explains many
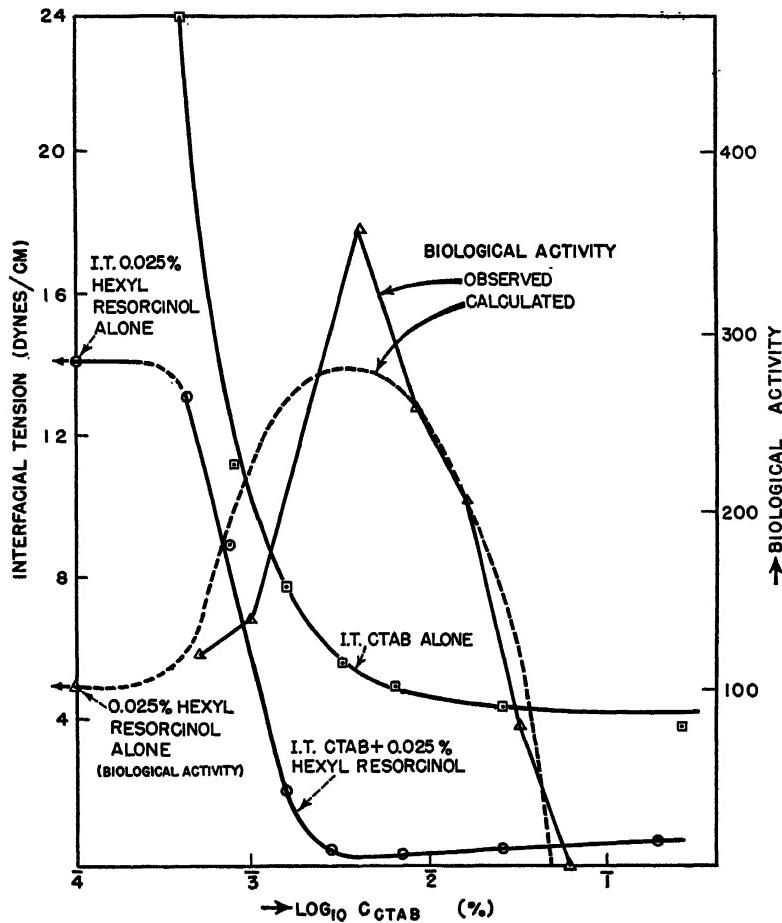
Figure 4. The biological and interfacial activities of hexyl resorcinol in the presence of the synthetic soap CTAB.

examples of a reduction in bacteriological activity observed in the presence of broth or agar.[24]

Biological Activity in Soap-Drug Mixtures

Most of the work referred to below concerns phenols in the presence of natural and synthetic soaps, but the concepts developed appear to be of general validity, and can be applied to other soap-drug systems.

In an *in vitro* examination of the effect of intestinal contents upon the rate of penetration of hexyl resorcinol into the pig roundworm (*Ascaris lumbricoides*), it was found that bile salts and sodium oleate had

marked effects. Synthetic soaps behaved similarly and a typical result with the cationic soap (CTAB*) is shown in Figure 4.[30] At low soap concentrations the rate of penetration is markedly increased, although the soap itself penetrates very little, if at all. At high soap concentrations the effect is reversed, and the penetration rate is ultimately reduced to zero. A precisely similar phenomenon is observed using bacteria as test organism.[31] The effects are clearly general ones and from measurements of the surface activity of the various mixtures a reasonable physicochemical interpretation can be given.

As pointed out above, structural features make soaps highly surface-active and also cause aggregation to micelles above a certain concentration. In the present systems, addition of soap in low concentration increases the surface activity of the mixture, thus increasing the adsorption on the surface of the organism and making it more permeable to the phenol. Up to the point of micelle formation the biological activity is thus increased. Once micelles start to form, however, they take up phenol from the solution and thus reduce the amount available for interaction with the organism. The biological activity thus falls to very low values if sufficient soap is added. The reduced uptake of phenol when micelles are present has recently been confirmed experimentally by direct measurement, using thick suspensions of bacteria.[28]

The Mechanism of Fat Absorption

The chemical and physical processes involved in the digestion of fat in the intestine have been the object of a number of recent investigations. In addition to chemical breakdown (hydrolysis) of the fat, emulsification was shown to be a most important factor. The way in which emulsification takes place spontaneously in the intestine (as it must do if rapid absorption is to occur) has been examined from the angle of surface chemistry, and this has shed a good deal of light upon the probable stabilizing agents.[32]

The mode of attack was to examine combinations of the principal intestinal constituents, such as triglyceride, monoglyceride, fatty acid, bile salts, etc., and see if any particular combination showed spontaneous emulsification at pH 6.5, the pH of the upper part of the small intestine. The only effective *in vitro* combination was the three-component-system fatty acid-monoglyceride-bile salt (taurocholate), and since these are normally present in adequate amounts, they are probably also responsible *in vivo*. The role of the bile salt in the interfacial film is probably to give a high charge to the emulsion particles, fatty acids being insufficiently ionized at an interface at pH 6.5. The way in which the

---

* CTAB is cetyl trimethylammonium bromide.

bile salt-monoglyceride-fatty acid mixture produces the requisite low interfacial tension ($<$ca. 1 dyne/cm for spontaneous emulsification) is clearly closely related to that in the mixed films mentioned earlier.

REFERENCES

1. Adam, N. K., "The Physics and Chemistry of Surfaces," 3rd ed., London, Oxford Univ. Press, 1941.
2. Alexander, A. E., and Johnson, P., "Colloid Science," London, Oxford Univ. Press, 1949.
3. Adam, N. K., *J. Chem. Soc.*, 164 (1933).
4. Alexander, A. E., *J. Chem. Soc.*, 1813 (1937).
5. Drake, N. L., and Wolfe, J. K., *J. Am. Chem. Soc.*, **62**, 3018 (1940).
6. Bilham, P., Jones, E. R. H., and Meakins, R. J., *J. Chem. Soc.*, 761 (1941).
7. ———, and Kon, G. A. R., *J. Chem. Soc.*, 552 (1941).
8. ———, ———, and Ross, W. C., *J. Chem. Soc.*, 35 (1942).
9. Anon., *J. Biol. Chem.*, **139**, 345 (1941); **143**, 171 (1942); **148**, 685 (1943); **173**, 383 (1948).
10. Crisp, D. J., *J. Colloid Sci.*, 1, 49, 161 (1946).
11. Glazer, J., Unpublished experiments.
12. Cumper, C. W., and Alexander, A. E., *Trans. Faraday Soc.*, 46, 243 (1950).
13. Alexander, A. E., *Proc. Roy. Soc. London*, **179A**, 470 (1942).
14. Bull, H. B., in "Advances in Protein Chemistry," eds. M. L. Anson and J. T. Edsall, Vol. III, p. 95, New York, Academic Press, Inc., 1947.
15. Allan, G., Unpublished experiments.
16. Schulman, J. H., and Rideal, E. K., *Proc. Roy. Soc. London*, **122B**, 29 (1937); *Nature*, **144**, 100 (1939).
17. Rothen, A., in "Advances in Protein Chemistry," eds. M. L. Anson and J. T. Edsall, Vol. III, p. 123, New York, Academic Press, Inc., 1947.
18. Alexander, A. E., in "Advances in Colloid Science," eds. H. Mark and E. Verwey, Vol. III, New York, Interscience Publications, Inc., in press.
19. Danielli, J. F., *Cold Spring Harbor Symposia Quant. Biol.*, **6**, 190 (1938).
20. Elkes, J. J., Frazer, A. C., Schulman, J. H., and Stewart, H. C., *Proc. Roy. Soc. London*, **184A**, 102 (1945).
21. Danielli, J. F., Danielli, M., and Marrack, J. R., *Brit. J. Exptl. Path.*, **19**, 393 (1938).
22. Mitchell, J. S., Rideal, E. K., *Proc. Roy. Soc. London*, **159A**, 206 (1947); **167**, 342 (1938).
23. Ferguson, J., *Proc. Roy. Soc. London*, **127B**, 387 (1939).
24. Trim, A. R., and Alexander, A. E., Soc. Exptl. Biol. discussion on "Antibiotics" in 1948 (1949, **III**, 111).
25. Alexander, A. E., Faraday Soc. discussion on "Recent Advances in Surface Chemistry" in 1947, Butterworths, p. 299, 1949.
26. Hartley, G. S., *Trans. Faraday Soc.*, **37**, 130 (1941).
27. Albert, A., Rubbo, S. D., Goldacre, R. J., Davey, M. E., and Stone, J. D., *Brit. J. Exptl. Path.*, **26**, 160 (1945).
28. Agar, A., and Alexander, A. E., *Trans. Faraday Soc.*, **45**, 528 (1949) and unpublished results.
29. McMullen, A. I., and Alexander, A. E., Faraday Soc. discussion on "Recent Advances in Surface Chemistry" in 1947, Butterworths, p. 309, 1949.
30. Alexander, A. E., and Trim, A. R., *Proc. Roy. Soc. London*, **133B**, 220 (1946).
31. Alexander, A. E., and Tomlinson, A. J. H., Faraday Soc. discussion on "Recent Advances in Surface Chemistry" in 1947, Butterworths, p. 317, 1949.
32. Frazer, A. C., Schulman, J. H., and Stewart, H. C., *J. Physiol.*, **103**, 306 (1944).

COMPLEXES OF IONS
WITH PROTEINS

Irving M. Klotz

Dept. of Chemistry, Northwestern University, Evanston, Ill.

INTERACTIONS BETWEEN PROTEINS and small ions have long been of interest to the biologist because of their significance in the function or dysfunction of many natural systems. Depending on its interests, however, each group of investigators has emphasized different aspects of these interactions. Thus, the physiologist has been concerned, to a large extent, with phenomena such as the excitatory (or inhibitory) effects of various anions or metallic cations on muscle activity. On the other hand, the

Figure 1. Schematic diagram of ion-protein complex.
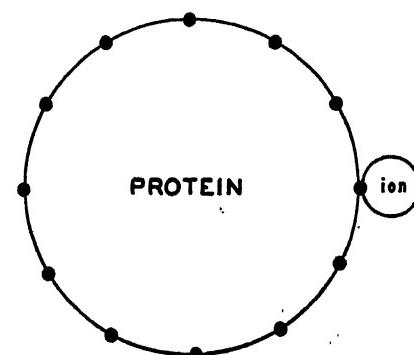
pharmacologist has been more interested in the binding of ions and molecules by serum proteins, for such information enables him to estimate effective drug levels in plasma, as well as to correlate excretion rates. The biochemist, in turn, has taken a special interest in interactions involving enzymes, since he recognizes that these macromolecules have special properties which frequently depend on their combination with organic or metal ions.

It has been realized for some time that an effective understanding of these interactions depends on an interpretation at the molecular level.

With most biological responses, however, there are several steps between the experimental observation and the initial ion-protein combination, which make it very difficult to establish the properties of the complex in an unequivocal fashion. On the other hand, with a crystalline, homogeneous protein, one is dealing with a much simpler system in which associations with small ions can be interpreted in a more straightforward fashion. It is for this reason that a great deal of effort' has been devoted recently to the elucidation of the nature of these complexes.

In an early stage of investigation, an ion-protein complex might be represented in a very schematic fashion by a diagram such as in Figure 1. Each constituent is indicated simply by a sphere, and possible sites of combination on the protein are represented by dots. There are numerous questions which one might raise in an attempt to add some detail to this model. In this paper, we shall consider four such questions:

(1) How many ions can be held by a single protein molecule?
(2) What is the strength of the bond between the protein and a specific ion, i.e., what is the energy of combination?
(3) What is the molecular and configurational nature of the site on the protein at which the ion is bound?
(4) What structural features in the ion favor combination?

Detailed answers to all of these questions cannot be given as yet, but a great deal of information has been accumulated in the past few years, and much of it can be used to obtain a clearer insight into the nature of ion-protein interactions.

Stoichiometric Relations

In any fundamental approach to the structure of an ionic complex, it is necessary to know first the stoichiometric composition of the aggregate. In the case of protein complexes, this problem has been obscured somewhat by the simultaneous existence of numerous stoichiometric ratios, the proportion of each one depending on the concentration of free, unbound ion in equilibrium with the complexes. For conciseness in representation of experimental data, it is convenient to plot the *average* number of ions per molecule of protein as a function of the concentration of the free ion. Thus in Figure 2, a graph of the data on copper-albumin complexes, one finds an average of seven cations on each protein molecule at a concentration of 10^{-3} molar ($\log = -3$) of cupric ion. It is necessary to recognize, however, that complexes exist in this same solution, which have fewer than seven cupric ions on each protein molecule, and that some are also present that have more than seven.

No method has yet been devised for measuring by a direct experiment the concentration of each type of complex in a given solution. Nevertheless, these concentrations can be calculated from theoretical equations, if

adequate data are available on the variation of the average number of bound ions with changes in concentration of free ion.

Proceeding to the problem of finding the maximum number of ions which can be held by a single protein molecule, we find that the answer can be obtained occasionally by direct experiment, but more often it requires some extrapolation procedure. In principle one would expect that the binding curve, as shown in Figure 2, should become horizontal as the concentration of free ion is increased. Such, indeed, is the case with
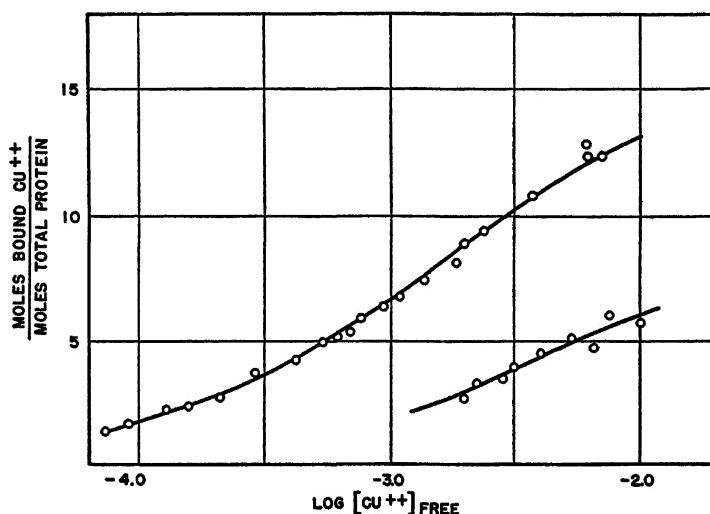
Figure 2. Binding of cupric ions by bovine serum albumin at 25°C. Upper curve at pH 4.83; lower curve at pH 4.00.

certain complexes between the protein lysozyme and some dye anions, where, at pH's near 7, the binding curve levels off when eleven ions are bound by each protein molecule. In most cases, however, the strength of binding is too weak to enable one to find this maximum directly. Thus in the case of copper–albumin complexes, it is difficult to tell from Figure 2 what the maximum number of bound copper ions can be. It is necessary, therefore, to resort to some extrapolation procedure.

Two different methods of extrapolation have been proposed in the literature. In one of these,[1] the data are plotted according to the equation

$$\frac{1}{r} = \frac{1}{kn}\frac{1}{c} + \frac{1}{n} \tag{1}$$

where r = moles bound ion per mole total protein
             c = concentration of free ion
             n = maximum number of ions which may be bound by a
                  single protein molecule
             k = binding constant

In a graph of $\frac{1}{r}$ versus $\frac{1}{c}$, the intercept on the ordinate is $\frac{1}{n}$; its reciprocal gives n. The second extrapolation procedure [2] makes use of the equation

$$\frac{r}{c} = kn - kr \tag{2}$$

In this case, r/c is plotted against r. The intercept on the ordinate is kn, and the slope of the line is k. Thus n can be determined from the ratio of intercept over slope.

For serum albumin the maximum number of ions combined with a single protein molecule has been reported for several different types of anion and cation. Some typical values have been assembled in Table 1. Even among the anions, no common value of n is apparent. Such a dependence of n on structure of the anion may reflect differences in configuration of amino acids around the sites at which binding occurs. On the other hand, a great deal of doubt has often been cast on the reliability of n values obtained by extrapolation.

TABLE 1. MAXIMUM NUMBER OF IONS HELD BY ONE MOLECULE
OF SERUM ALBUMIN*

Ion	Maximum Number Bound	Reference
Chloride	6	(3)
Dodecyl sulfate	110	(4)
Dodecyl sulfate	14	(5)
Phenyl butyrate	24	(6)
o-Nitrophenolate	6	(6)
m-Nitrophenolate	24	(6)
p-Nitrophenolate	25	(6)
Methyl orange	22	(1)
Cupric	16	(7)

* No distinction is made in this table between albumins of different origin, that is, whether human, bovine or equine.

Thus, the significance of differences such as those reported in Table 1, is still open to question.

Energy of Binding

To evaluate binding energies one must recognize again that multiple complexes are formed in ion-protein aggregates. Thus it is necessary to represent the equilibria involved by a set of equations such as

$$P + I = PI$$
$$PI + I = PI_2 \tag{3}$$
$$\cdots \cdots \cdots$$
$$PI_{i-1} + I = PI_i$$

where P indicates the protein molecule and I the small ion. For each of these equilibria, one can then formulate an equilibrium constant which in general terms may be written

$$\frac{(PI_i)}{(PI_{i-1})(I)} = k_i \qquad (4)$$

It is evident, that the extent of binding of a particular ion, I, by a specified protein would depend on the magnitude of the equilibrium constants. The affinity of the protein for the ion thus can be measured quantitatively in terms of k_i, and for expression in thermodynamic terms, the free energy of binding can be evaluated from the relation:

$$\Delta F_i^\circ = -RT \ln k_i \qquad (5)$$

As usual, R represents the gas constant and T the absolute temperature.

The actual evaluation of ΔF_1°'s from experimental data involves relatively extensive computations, in which statistical, electrostatic, and perhaps other factors need be considered.[1, 2, 5] For our present purpose, however, to give an indication of the order of magnitude of the energy of binding, it will suffice to restrict our discussion to ΔF_1°, the free-energy change for the formation of the first complex. Some typical binding energies are listed in Table 2. All values are in the neighborhood of 5 kilocalories/mole. It is obvious, then, that the bonds formed are loose ones and hence not of the strong covalent type. Relationships between binding energy and structure of the small ion or molecule will be considered in one of the following sections.

Equal in interest to the free-energy of binding, are the associated thermodynamic quantities, the entropy (ΔS) and enthalpy (ΔH) of binding.* Quantitative data which permit the calculation of all three of these quantities have been obtained for only a very few ion–protein complexes.[5, 7, 11] Nevertheless, a common feature is discernible already. All of the ions examined (five anions, one cation) show a small enthalpy of binding, but a relatively large and *positive* entropy of binding. For an association process, such as is represented by equation (3), an increase in entropy is rather surprising. Even more unexpected, however, is the observation that the increased binding energy for larger ions is not caused by any substantial increase in ΔH, but almost entirely by larger positive ΔS values.*

---

*According to the theory of thermodynamics, the magnitude of the free-energy change (ΔF) for a given chemical reaction depends on two related thermodynamic quantities, the enthalpy change (ΔH) and the entropy change (ΔS). Specifically, $\Delta F = \Delta H - T\Delta S$, for a reaction at a constant temperature, T. The ΔH term may be identified directly with the heat absorbed in a chemical reaction under normal conditions. Thus, the more heat *evolved*, the more negative will be ΔH, and hence the more negative the contribution to ΔF. The entropy term, ΔS, is interpreted most conveniently on a molecular basis. When a molecular system changes to a less orderly arrangement, S increases and ΔS is positive;

A full discussion of the significance of these thermodynamic quantities would not be appropriate in a brief review. Nevertheless, in connection with subsequent discussions on the nature of the ion–protein bond, it is pertinent to point out that small ΔH values and large positive ΔS values for an association process are characteristic of electrostatic reactions. This behavior has been found experimentally to be typical of reactions such as acid–base equilibria. Large ΔS values and small ΔH's for electrostatic combinations would be predicted also from theoretical considerations.

Nature of Binding Site on Protein

In considering the characteristics of the locus of attachment on the protein, we must treat complexes with anions separately from those with cations.

TABLE 2. BINDING ENERGIES FOR COMPLEXES WITH SERUM ALBUMIN

Ion or Molecule	Binding Energy, $(-\Delta F_1^0)$ calories/mole	Reference
o-Nitrophenolate	5565	(6)
m-Nitrophenolate	4965	(6)
p-Nitrophenolate	4865	(6)
Salicylate	4780	(8)
Methyl orange	5960	(1)
Adenine	3080	(9)
Adenosine	3370	(9)
Adenylate	3810	(9)
Sulfanilamide	3740	(10)
Cupric	5180	(7)

Much evidence has now been accumulated to indicate that anions are held by the protein largely by the electrostatic forces of positively-charged loci of one of the basic amino-acid residues (Figure 3). As has been pointed out in the preceding section, the thermodynamic quantities associated with complex-formation exhibit values characteristic of an electrostatic reaction. In addition it has been observed [12, 13] that the extent of binding of anions drops very rapidly in solutions of pH 11 and above, i.e., in solutions where the cationic amino acids lose their positive charge.

vice versa, when a system becomes rearranged into a more orderly configuration, S decreases and ΔS is negative. As is evident from the equation $\Delta F = \Delta H - T\Delta S$, if ΔS is positive, ΔF becomes more negative.

Thus, the more negative the value of ΔH and the more positive the value of ΔS, the greater will be the binding energy; for the binding energy is defined as $-\Delta F$, since the greater the *drop* in F, the more stable is the complex formed in the reaction under consideration. In other words, the greater the heat *evolved* and the more disarranged the system formed at the end of a reaction, the more stable is the resultant product.

Even more striking evidence has been obtained by an examination of the binding affinity of proteins chemically modified in a controlled manner. It is possible to acetylate selectively the ε-amino groups of the lysine residues of a protein under such mild conditions that the framework of the macromolecule remains intact. The net result of this modi-
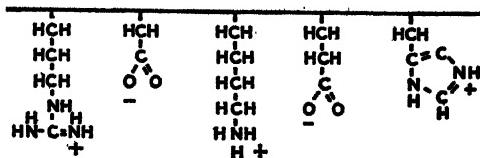
Figure 3. Schematic diagram of protein molecule.

fication, then, is the elimination of a positively-charged locus on the protein:

$$\text{P}\!-\!\text{NH}_3^+ + (\text{CH}_3\text{CO})_2\text{O} = \text{P}\!-\!\text{NHCOCH}_3 + \text{CH}_3\text{COOH} + \text{H}^+ \qquad (6)$$

In consequence, one would expect a substantial drop in binding ability, and such indeed is the case.[14] On the other hand, it is possible also to convert the ε-amino group of the lysine residues to guanidine groups by the reaction:

$$\text{P}\!-\!\overset{+}{\text{N}}\text{H}_3 + \text{CH}_3\text{OC}\!=\!\text{NH} = \text{P}\!-\!\text{NHC}\!=\!\overset{+}{\text{N}}\text{H}_2 + \text{CH}_3\text{OH} \qquad (7)$$
$$\qquad\qquad\qquad | \qquad\qquad\qquad\qquad |$$
$$\qquad\qquad\quad \text{NH}_2 \qquad\qquad\qquad\quad \text{NH}_2$$
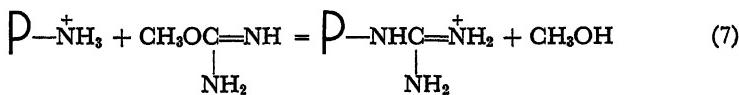
In this case, the nature of the site is altered, but the positive charge is retained. As would be predicted, then, the ability to bind anions is retained.[14]

While the presence of a positively-charged locus is apparently a necessary condition for the protein to bind anions, it is not a sufficient condition. Thus with a small anion, such as methyl orange, only two proteins, serum albumin and β-lactoglobulin show appreciable binding affinity at physiological pH's. On the other hand many other proteins, for example serum γ-globulin, have practically as many cationic, amino-acid residues as does β-lactoglobulin, and yet do not form anion complexes. The basis of this difference in behavior is not yet certain, although it probably involves the configuration of the amino acids around the cationic locus. It has been suggested recently[14] that hydroxy and dicarboxylic amino-acid residues within the protein in the neighborhood of the cationic group may compete with the small anion attempting to come in from the outside. On the basis of this picture it has been possible to establish a quantitative *binding index*, based on the distribution in con-

tent of amino acids in a protein, which predicts the relative binding ability of the protein. This index fits the observations accumulated so far, but a rigorous test of its reliability must await further investigations.

Turning to complexes of proteins with a metallic cation such as copper, we might be inclined to attribute binding properties to the presence of negatively-charged carboxylate groups on the macromolecule (Figure 3). The presence of carboxylate residues does indeed seem to be a necessary condition, at least for binding in the neighborhood of pH 5;
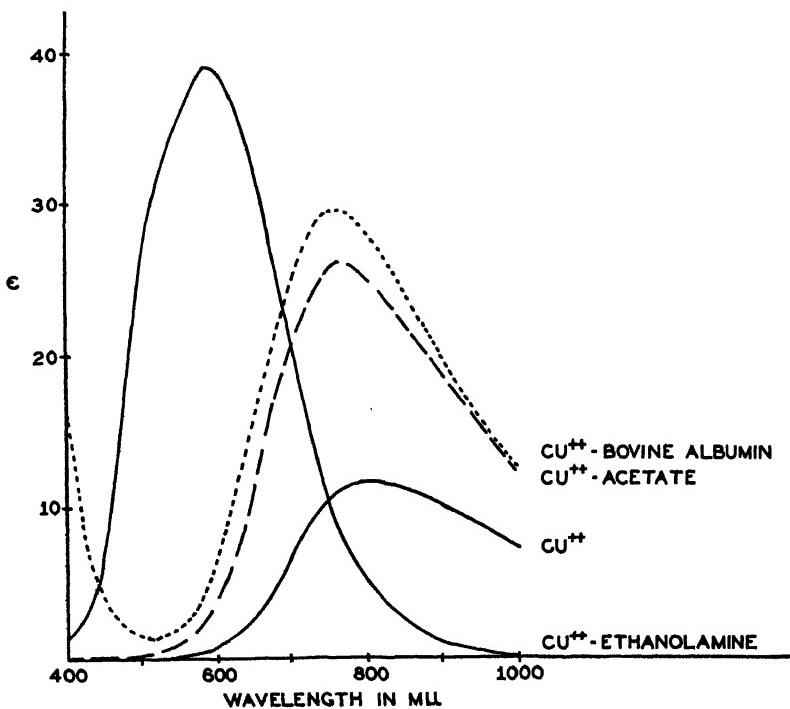
Figure 4. Absorption spectra of some copper complexes in aqueous solution near pH 5.

for if the pH is lowered, and the COO⁻ groups converted to uncharged COOH residues, the extent of binding of Cu⁺⁺ drops rapidly.[7]

Perhaps even more convincing, however, are some of the spectrophotometric observations illustrated in Figure 4. The optical absorption of Cu⁺⁺ depends on its environment, as one would expect. When complexed with an amine group, the absorption peak of Cu⁺⁺ is generally in the neighborhood of 600 mμ, as is shown for the ethanolamine case in Figure 4. On the other hand, with COO⁻ groups, such as acetate, the Cu⁺⁺ complex shows a maximum absorption slightly below 800mμ. It is of interest, then, to notice that the copper-albumin complex (near pH 5)

has a peak almost identical in wavelength with that of copper acetate. A similar spectrum is obtained with the copper-β-lactoglobulin complex. The conclusion seems inescapable, then, that COO$^-$ groups on the proteins are intimately involved in the bond with the metallic cation.

Effect of Structure of Anion

Because of its possible significance in connection with biological specifications, a great deal of attention has been paid to the effect of structural
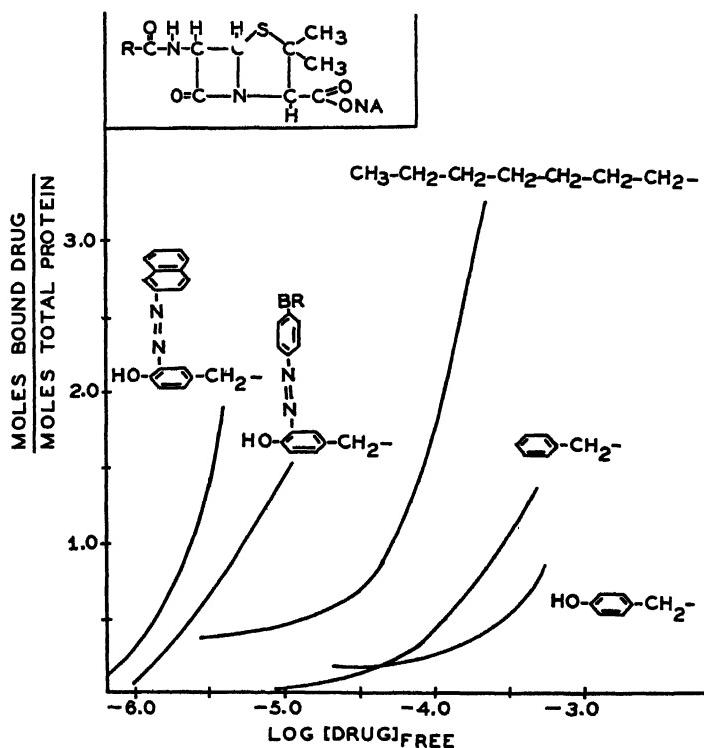
Figure 5. Binding of penicillins of increasing molecular weight by serum albumin.

variation on the affinity of anions for serum albumin. It should be emphasized at the outset, however, that this protein shows a very nonspecific character in its ability to bind a vast variety of anions. Nevertheless, certain anions are bound more strongly than others, and an examination of the structural basis of these differences should lead to some principles of general applicability.

Numerous comparisons have been made within pairs of molecules of practically identical structure, but differing in the presence or absence

of a negative charge. Invariably the anion has been bound much more strongly. Thus albumin has a much greater affinity for [15]

$$H_2N\underset{}{\bigcirc}SO_3^- \quad \text{than for} \quad H_2N\underset{}{\bigcirc}SO_2NH_2$$
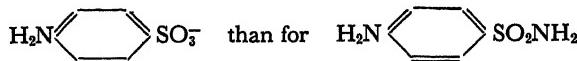
Similarly, $CH_3(CH_2)_4COO^-$ is bound much more strongly than $CH_3(CH_2)_4CONH_2$.[16]

Within a series of anions of similar base structure but increasing molecular weight, binding affinity increases with size. This behavior is illustrated in Figure 5 for penicillin-albumin complexes. Similar behavior has been observed in the series of aliphatic acids [16] and has been pointed out even earlier [17] in connection with titration curves for proteins. The principle of increased binding with increased size is, in fact, implicit in the observation of Sørensen [18] that the protein error with indicators is generally smaller with indicators of simpler constitution.

Some interesting observations have also been made recently in connection with the relative affinities of isomeric anions. Thus,

$$\underset{\text{OH}}{\overset{CH_2COO^-}{\bigcirc}}$$
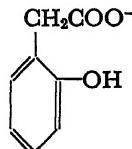
has been found [19] to be much more strongly bound than

$$\underset{\text{OH}}{\overset{CH_2COO^-}{\bigcirc}}$$
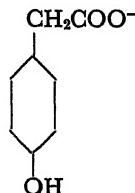
Similarly albumin shows a stronger affinity for

$$\underset{}{\overset{SO_3^-}{\bigcirc}}-NH_2$$

than for

$$\underset{NH_2}{\overset{SO_3^-}{\bigcirc}}$$
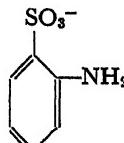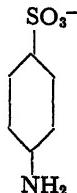
It seems likely that the ortho compounds form intramolecular hydrogen bonds which decrease the affinity of the aqueous solvent for the anion and hence make it easier for the protein to hold on to the small ion.

Conclusions

The foregoing discussion represents a rather cursory survey of some of the characteristics of ion-protein complexes. Most of the work described has been carried out with the serum proteins, and hence most of the conclusions derived, of general biological import, have been in connection with the physiological functions of plasma proteins. Thus it has been emphasized [20] that these proteins may play a very important role in the transport of a variety of ions and molecules within the organism. Similarly it has been pointed out that serum albumin may act as a *concentration buffer* by tying up small ions so that they are not excreted too rapidly by the kidneys.

The study of ion-protein complexes is also beginning to give some information with more fundamental implications. Even with serum albumin, structural specificities, of which those described are but a few, are becoming evident among the groups of anions being studied. These interactions, sooner or later, must give some direct indications of the basis of enzyme specificities.

Perhaps the most significant contribution which can come from studies of these complexes is in connection with the structure of proteins. Even from the limited investigations carried out so far, an inkling is being obtained of the configurational patterns at the surface of the protein molecule. Further studies will furnish more detailed knowledge of these configurational arrangements and thus will provide a foundation for a rational approach to many of the puzzling problems of biology.

REFERENCES

1. Klotz, I. M., Walker, F. M., and Pivan, R. B., *J. Am. Chem. Soc.*, **68**, 1486 (1946).
2. Scatchard, G., *Ann. N. Y. Acad. Sci.*, **51**, 660 (1949).
3. ———, Batchelder, A. C., and Brown, A., *J. Am. Chem. Soc.*, **68**, 2320 (1946).
4. Putnam, F. W., and Neurath, H., *J. Biol Chem.*, **159**, 195 (1945).
5. Karush, F., and Sonenberg, M., *J. Am. Chem. Soc.*, **71**, 1369 (1949).
6. Teresi, J. D., and Luck, J. M., *J. Biol. Chem.*, **174**, 635 (1948).
7. Klotz, I. M., and Curme, H. G., *J. Am. Chem. Soc.*, **70**, 939 (1948).
8. ———, Triwush, H., and Walker, F. M., *J. Am. Chem. Soc.*, **70**, 2935 (1948).
9. ———, and Urquhart, J. M., *J. Biol. Chem.*, **173**, 21 (1948).
10. ———, and Walker, F. M., *J. Am. Chem. Soc.*, **70**, 943 (1948).
11. ———, and Urquhart, J. M., *J. Am. Chem. Soc.*, **71**, 847 (1949).
12. Boyer, P. D., Ballou, G. A., and Luck, J. M., *J. Biol. Chem.*, **167**, 407 (1947).
13. Klotz, I. M., and Walker, F. M., *J. Am. Chem. Soc.*, **69**, 1609 (1947).
14. ———, and Urquhart, J. M., *J. Am. Chem. Soc.*, **71**, 1597 (1949).
15. ———, *J. Am. Chem. Soc.*, **68**, 2299 (1946).

16. Boyer, P. D., Lum, F. G., Ballou, G. A., Luck, J. M., and Rice, R. G., *J. Bic Chem.,* **162,** 181 (1946).
17. Steinhardt, J., *Ann. N. Y. Acad. Sci.,* **41,** 287 (1941).
18. Sørensen, S. P. L., *Biochem. Z.,* **21,** 131 (1909).
19. Luck, J. M., and Schmit, A. S., *Stanford Med. Bull.,* **6,** 133 (1948).
20. Davis, B. D., *Am. Scientist,* **34,** 611 (1946).

CATALYSTS AND NEOCATALYSTS IN
BIOLOGY AND MEDICINE

Jerome Alexander

THE CHEMICAL CHANGES resulting from the activities of any catalyst—
and all enzymes are biocatalysts—take place at, or very close to, the
catalyst surface whose specific physicochemical configuration is a domi-
nant factor in controlling what happens there. The nature and rate of
the catalytically-directed chemical changes may vary with the local tem-
perature, the pressure, the relative proportions of reactants and of other
substances present, and with the speed of removal of new products
formed. With most organisms, the presentation of reactants to active
catalyst areas and the removal therefrom of reaction products and of
substances influencing reaction are controlled by differential diffusion
through tissues, septa, and cellular protoplasm; by the massive distribu-
tion of these products by circulating fluids (e.g., blood, lymph); and by
their removal by selective adsorption, polymerization, chemical reaction,
or utilization by other catalysts.

Tempo in Biological Reactions

For the most part, biocatalysts exist within cells or body fluids in col-
loidal dispersion, whether free (enzymes) or else fastened to or adsorbed
on tissue elements. In discussing the zone of maximum or optimum
colloidality in Volume I of this series,* the writer referred to the control
exerted by degree of dispersion on the velocity of chemical reaction as
follows:

"A most striking example of optimum dispersion is found in living
matter. Figuratively speaking, if all the chemical substances comprising
our organism were in true or crystalloid dispersion, reactions would
proceed so rapidly that we would, so to say, live ten years in ten minutes.
On the other hand, if coarse dispersion prevailed, it would take ten

* "Colloid Chemistry, Theoretical and Applied," Ed. J. Alexander, Vol. I, pp. 25-26,
New York, Reinhold Publishing Corp., 1926.

237

years to live ten minutes. Every organism is dependent upon the co
ordination of its chemical reactions *in point of time,* and this leisurely
procedure depends largely on *degree of dispersion,* which keeps chemical
reaction velocities within certain speed limits through its regulation of
free surface and kinetic velocity. Life lies between lysis and coagulation.
The colloidal zone is, as it were, a vital metronome tolling off the tempo
of life."

Besides controlling the *tempo* of chemical changes in organisms, the
biocatalysts also control the *specificity* of these changes. They are not
the sole controlling factors, because not only must essential atoms and
molecules be present in the food and milieu, but these must also reach
the biocatalyst areas by diffusion and circulation at suitable speeds, and
reaction products must likewise be disposed of by local utilization or by
removal.

Heat in Biological Reactions

Since the heat liberated at or near a biocatalyst surface may inactivate
or destroy the heat-sensitive enzymes, operative natural processes must
be provided to dispose of this energy. If the entire heat of oxidation of
dextrose in the body (700 Cal per mole) *were* suddenly released, the
enzyme system would be inactivated or even charred. However, bio-
logical catalysts are protected against such an occurrence. Thus, the
heat set free by the bacterial oxidation of organic substances is so slowly
evolved that farmers can utilize "steaming" manure in hot-beds and in
trenches safeguarding buildings in the winter. Similarly, industrial cata-
lysts may be protected. For example, the use of iron tubes filled with
mercury (which boils at 357° C) is described in a patent,[1] to prevent
catalyst fusion caused by the high reaction heat which is released when
phthalic anhydride is produced by passing a mixture of air and naph-
thalene vapor over a vanadium pentoxide catalyst. To cite an applica-
tion of "cooling protection" drawn from every-day experience: a grind-
stone will quickly ruin the temper of an axe if cooling water is not
properly supplied. Pliny [2] observed that an oil-whetstone gives a keener
edge than a water-whetstone. Lubrication by the oil prevents the local
frictional development of too high a temperature, and is more effective
than water in maintaining the temper of the cutting edge.

Three factors dominate the protection of the organism against damage
by biochemically-liberated heat:

(1) The reactions usually take place in a *stepwise* manner; there are
a series of chemical changes, each of which is mediated by its own spe-
cific biocatalyst.

(2) Each intermediate substance must move from the place of its

production to the next catalyst area in the series; this demands *time.* The various reactions thus take place *intermittently* and in a *spotwise* manner, the increments of heat being liberated in variable "quanta" at different localities.

(3) The several chemical changes take place in *aqueous* media, so that water with its high specific heat not only prevents too great an increase in local temperature, but also conserves heat locally. In the case of large organisms, the internally-liberated heat is carried to exterior surfaces by circulating fluids and is there radiated away. It seems likely that the sufferings of the packed-in victims of the "Black Hole" of Calcutta were largely caused by the inability to radiate away body heat, and not exclusively to lack of oxygen and excess of carbon dioxide.

The importance of phosphate in biological oxidation, observed by Pasteur (1860) and confirmed by Young and Harden (1905), is being intensively investigated by the Cori, Lipmann, Meyerhof, and Kalkar schools, among others. Summing up recent views Brody [3] stated:

". . . . some phosphate esters serve as temporary biologic energy reservoirs, analogous to charged batteries. Thus, according to Cori, the synthesis of 6 molecules of glucose phosphate is coupled, or associated, with the oxidation of one molecule of glucose. A mol of glucose phosphate, therefore, has a labile energy increment which, depending on the energetic efficiency of the process, may be as high as 115 Cal (one-sixth of about 700 Cal, the free energy of glucose). This is, presumably, what Lipmann refers to as phosphate-ester bond energy, the main form or source of anaerobic energy as illustrated by the reaction:

$$\textit{Glycerophosphate} + H_2O \rightleftharpoons \textit{Glycerol} + \textit{Phosphate} + \textit{Energy}$$

. . . The phosphate group also catalyzes the oxidation and transport of fats. Other inorganic elements may participate in the oxidation of fat and perhaps its transformation to carbohydrate." Most electron donors (metabolites) must be phosphorylated as the preliminary step, for phosphorylation and oxidation are indeed *coupled* reactions.

What is the mechanism whereby the so-called "energy-rich" phosphates are able to deliver their energy to facilitate chemical change? While it is conceivable that in some cases energy may be transferred without loss from one molecular grouping to another, it seems reasonable to believe that despite the cooling effect of aqueous media, the heat liberated by chemical changes may increase the kinetic motion, resonance, or intraparticulate activity of *some* molecules, so as to give them momentarily what corresponds to the *effect* of a relatively high temperature. A molecule or a specific portion of a large molecule (including the catalyst) could thus become highly reactive, and the thermal energy could be locally utilized without any dangerous increase in local temperature.

In considering the significance of coupled reactions for the enzymic hydrolysis and synthesis of proteins, Bergmann and Fruton [4] observe that the thermodynamic data alone indicate merely the approximate amount of energy needed to make the system operative. "We must therefore look for the specific physical and chemical mechanisms which make possible the synthesis of peptide bonds." The work of Schoenheimer "led him to the view that in the dynamic equilibrium between proteins and amino acids in the tissues, peptide bonds are continually being broken and re-formed under the catalytic influence of the tissue enzymes." The violent Brownian motion seen in the ultramicroscope gives some indication of the much more violent but unseen molecular agitation at lower levels of structure, which is greatly increased by locally-liberated heat.

Efficiency in Biocatalysis

A standard cadmium storage battery or cell, if slowly charged and discharged, may approach 100 per cent efficiency; but for steam engines the efficiency is about 20 per cent, for gas engines about 25 per cent, and for Diesel engines about 40 per cent of the fuel energy (although the results in many cases may fall far below these figures). Much of the energy is radiated away as heat, as is also the case with animals. However, animals utilize much of the energy in their foods in chemical processes involving syntheses which require heat and lead to growth, reproduction, or maintenance. While the efficiency of some of these synthetic processes may be low, the waste heat may be needed to initiate or speed up other processes, as well as to maintain body temperature, so that the overall utilization of the food energy is high. Rahn [5] estimates that one Calorie in food liberates the following amounts of heat (in Calories) in various organisms: pig, 0.2 to 0.4; trout, 0.18 to 0.31; cockroach, 0.34 to 0.35; mold, 0.58 to 0.70; colon bacillus, 0.13 to 0.24; pseudomonades, 0.21 to 0.22. *Nitrobacter,* in using the energy derived from the oxidation of nitrite to nitrate, to assimilate carbon from carbon dioxide, shows an efficiency of only about 6 per cent, and autotrophic bacteria, in producing ammonia from atmospheric nitrogen, approximate only 3 per cent efficiency. Much depends upon the criteria used to calculate "efficiency"; agriculturists may estimate it in terms of marketable products—wool, meat, fat, eggs, milk, skins or hides, manure. As the term "horse-power" indicates, muscular energy may also be important. If all factors are considered, the gradual liberation of energy by serial catalyst changes will add up to a rather high efficiency figure.

The large-scale industrial production of many organic and inorganic compounds is being carried out and extended because both the active catalyst surfaces and the operating conditions are closely controlled. De-

spite the extreme sensitiveness of biocatalysts (enzymes), many microorganisms (yeasts, molds, bacteria) are successfully used in producing alcohol, citric acid, lactic acid, penicillin, etc. With the higher plants and animals, the number of catalyzed reactions regularly carried on is enormous. The fact that the ship of life can steer a more or less definite course in the various and variable conditions it must meet, indicates how generally dependable are the catalytic mechanisms not only in carrying on the chemical processes essential to the existence and maintenance of the organism, but also in adapting themselves to new conditions which continually arise in the life of individuals and of the race. A basic factor in the adaptability of living things is their power to develop novel catalyst surfaces, or *neocatalysts*, which, by altering the nature and/or rate of the chemcial output, give a new direction to the course of life.

How Do Neocatalysts Arise?

Neocatalysts may be formed *de novo* by the aggregation of subunits in such a manner that there is formed a new surface mosaic which is catalytically functional in a novel way. For example, thiamin is first converted into a pyrophosphate to act as an activating or prosthetic group which, on being taken up by a suitable carrier, forms the enzyme cocarboxylase. Neocatalysts may also arise by the alteration of an existing catalyst surface, a process which the author terms *modification*. One cannot be too dogmatic about this terminology, for the creation of a catalyst surface *de novo* is simply a limiting case, where the catalytic activity of the earlier surfaces is zero.

The main ways by which catalyst surfaces are modified are: (1) by physicochemical readjustment which does not involve the addition of new material to the catalyst mass (e.g., electroversion,[6] some types of gene mutation); and (2) by additions to, or subtractions from, the catalyst surface.

Novel prosthetic groups and carriers may appear in an organism in quite a number of different ways: e.g., from foods (vitamins); as a result of a gene mutation; because they are set free by elution, deflocculation, or chemical reaction; from the invasion of a pathogenic or a helpful organism, or through a symbiosis.* Svedberg has shown that aggregation and deflocculation may follow slight ionic changes, or the presence of minute amounts of specific substances; and catalytic potency may appear or disappear with these changes. Often the insertion of a single metal

---

* For example, the mold and the alga, which constitute a lichen, cannot now survive independently in nature, although each can be cultivated separately in the laboratory. . . . Prof. L. R. Cleveland killed the endameba "messmate" of termites by increased oxygen pressure, and the termites died because they cannot digest wood, but simply absorb products of the messmate's catalysts.

atom into a surface may establish, intensify, or alter its catalytic action. Thus, only iron phthalocyanine exhibits marked catalase activity; the cobalt and nickel compounds act feebly.

Promotors, as they are called in industry, are substances introduced into a catalyst surface to favor the formation of desired products. The suppression of undesirable reactions may be of equal importance, and may be effected by "beneficial poisoning" of limited catalyst areas; for as Taylor showed (1925), catalysts may have more than one specifically-active facet. Thus, nickel, poisoned by thiophene for the hydrogenation of benzene, will hydrogenate ethylene; and even when poisoned for both of these reactions, will still serve for the reduction of nitrobenzene. On the other hand, a "poisoned" catalyst surface may be liberated for activity by the removal of inhibitors; e.g., dimercaptopropanol, known as BAL (British anti-lewisite),

$$\begin{array}{ccc} OH & SH & SH \\ C & - C & - C \\ H_2 & H & H_2 \end{array}$$
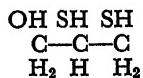
serves as an antidote in poisoning by arsenic or mercury, probably because its active SH–groups remove the metal atoms from the SH–groups of the poisoned catalysts. Slight changes in catalyst surfaces may be most significant, giving rise to a neocatalyst which can introduce and perpetuate a new chemical output.

Let us now consider some of the basic phenomena of biology and medicine in the light of the comparatively recent notion of neocatalysts.

Self-Saving Catalysts in Immunology and Serology

Immunity, in its medical sense, means successful resistance to disease, and is commonly brought about by specific substances which "kill" or inactivate the catalysts of invading pathogens, or nullify the action of harmful substances produced by pathogens. These protective or neutralizing *antibodies* are often found in the blood serum after a natural or an induced infection (active immunity), or they may be formed in the blood of another animal and then separated and injected to produce immunity (passive immunity). Antibodies may enter the blood of the new-born from the mother and persist for a time. Serology, the science of blood sera, is a term more limited than immunology, although sometimes erroneously used as a synonym.

Conditions may affect immunity. For example, the immunity of chickens to anthrax is attributed to their high body temperature (about 103°F); for a chicken chilled in ice water and then inoculated with anthrax develops the disease.

In order to be effective, antibodies must be able to persist in sufficient

quantity and to reach the pathogen or its products in good time. We have much to learn about the various factors which produce or prevent immunity in different animals for specific pathogens.

The Mechanism of Antibody Formation. About 1925, the work of the Bragg's on crystal structure, and of Langmuir and others on specific adsorption, led the author to visualize the formation of specific antibodies against specific antigen particles. These antigen particles made up a "master mold" capable of shaping great numbers of particles (not necessarily single molecules) having a surface pattern specifically the reverse of that of the mold. In discussing this notion privately with scientists of all kinds, and also publically in lectures before the American Chemical Society and elsewhere, the writer used a coin to represent the specific antigen surface and showed that on pressing a piece of tinfoil
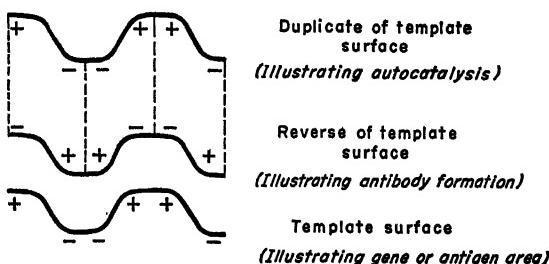
Figure 1. Diagrammatic section of a specific surface, showing the complementary relation between reproductive catalysis (autocatalysis) and antibody formation.

against it, the near surface of the foil acquired the precise reverse pattern. The upper surface of the foil made a duplicate of the coin surface, a model for reproduction at the molecular or near-molecular level of structure (Figure 1).

Following five or six years of open discussion which revealed no valid objection to this notion and no imaginable alternative, a brief paper was prepared and sent first to one, and then to another American biological journal; both rejected it, one highly-placed referee writing that "there are an infinite number of similar speculations possible." The paper was therefore sent to *Protoplasma,* where it was published in October, 1931. Boyd, in his book, "Fundamentals of Immunology," states: "The first theory of antibody formation which still seems at all probable today is that put forward independently by Breinl and Haurowitz,[7] Alexander,[8] and Mudd [9] (see Hooker and Boyd[10])." A copy of the *Protoplasma* paper was sent to Miall, editor of *Chemistry and Industry* (London), who wrote an editorial [11] in 1932, in which he used the apt term "template" to describe the master-mold function of the antigen (see also Alexander [12]).

Since then, many others have used the term "template" in this connection, as well as the tinfoil model, apparently in ignorance of the origin of these ideas.

Without going into further detail, it is evident that we have here a mechanism whereby a foreign substance introduced into the body *might* direct the formation of particles of opposite contour. Where this *does* occur, *and* the specificity of the reverse surface is maintained, *and* the blood serum can be shown to carry specific antibodies, the foreign substance is called an *antigen*. In effect, the antigen acts like the prosthetic group of a catalyst. For this reason, Sevag[13] called his book on immunology "Immuno-catalysis."

It is obvious that the ability to form neocatalysts capable of producing lasting and efficient antibodies constitutes an important factor in evolutionary survival; for organisms are thereby protected against pathogens which, apart from producing harmful substances, may utilize or destroy molecules essential to the welfare of the host. Such neocatalysts may, therefore, be termed "self-saving catalysts."

Embryology—the Catalyst Entelechy in Morphogenesis and Differentiation

It was only about 125 years ago (1827) that von Baer first observed a mammalian egg, that of a dog. The eggs of rodents (rats, mice, guinea pigs) are about 70 to 85 microns in diameter; those of horses, dogs, sheep, goats, pigs, whales, monkeys, apes and man are about 140 microns. Yet these almost microscopic spherical units contain determinants which enable them when fertilized by an appropriate sperm, to develop under suitable conditions into highly-specific though widely-differing individuals.

In attempting to account for this regularly-occurring miracle, let us estimate what the tiny fertilized ovum (called the zygote) means in terms of molecules, figuring molecular dimensions at 0.2 to 5.0 millimicrons $(m\mu)$ and the zygote as a sphere whose diameter is 100 microns, that is, 100,000 $m\mu$. The volume of this sphere $(v = \frac{1}{3} \cdot r^3)$ is approximately 262×10^{12} $m\mu^3$, that is, 262-million-million cubic millimicrons. Since a major portion of the molecules in the zygote are small ones (water, salts, dextrose), there is ample room not only for the chromosomes (gene strings), enzymes and other inclusions, but also for many many millions of highly specific molecules supplied mainly by the cytoplasm of the ovum, but in part by the sperm of the father. It is quite possible, therefore, for the zygote initially to contain, in addition to the Mendelian genes and many specific enzymes, enormous numbers of specific molecules and ions capable of serving as modifiers of existing catalysts, or as pros-

thetic groups or carriers to form additional neocatalysts as development proceeds. The introduction of other molecules from food, from the milieu, and from symbionts or their activities, offers many possibilities for the formation of still other neocatalysts, with new chemical ambits, leading on and on.

The major portion of our food molecules are oxidized or otherwise disposed of as "fuel," so to say, in the fires of life's enzymic catalytic processes; but many remain for a longer or shorter time in body fluids, cells, and other structures, and some, especially vitamins, iodine, iron, and copper, take part in the formation of specific catalysts. The body is continually wearing out changing and replacing the various particulate units of its structure, fluids, and operative mechanisms. Experiments with tagged atoms indicate that there may be a slow replacement of atoms in the constituent chemical compounds of bones and teeth. Most molecules are transient guests in the hotel of our bodies, and in the course of our development anatomical structures appear, only to undergo profound change, reconstruction, or total or partial obliteration. Sometimes residues remain to remind us, e.g., the lachrimal caruncle, and the appendix. All this is well known to anatomists and to comparative embryologists, and led von Baer to the view that each individual in the course of its development passes through many of the type forms of its phylogenetic group. This process is termed *recapitulation*.

Embryologists have long recognized the fact that specific substances, small in mass but potent in effect, dominate the development of the zygote. The term *organizer*, introduced by von Spemann, is defined by Needham [14] as ". . . A living part of an embryo which exerts a morphogenetic stimulus upon another part or parts, bringing about their determination and the following histological and morphological differentiation. The organizer which acts first in development is known as the *Primary* or *First-grade* Organizer those acting at successive stages are known as *Secondary (Second-grade)* and *Tertiary (Third-grade)* Organizers. The appearance of the latter is generally dependent upon the prior function of the primary organizer, which, therefore, *seems* to be more manifold in its inductive effects. The chemical substance emitted by an organizer is called an *Evocator*."

The term *"induction"* [15] is thus defined as ". . . The morphogenetic effect brought about by an organizer, inductor, evocator, etc., acting on competent tissue." And "competence" [16] is described as "The state of reactivity of a part of an embryo, enabling it to react to a given morphogenetic stimulus by determination and differentiation in a given direction. An embryonic cell develops as many competences as it has prospective potencies or possible morphogenetic fates. Generally speaking, determination to proceed to one morphogenetic destination cancels the

other competences previously possessed by the cell. Competences appear
in a succession at definite stages of development and later disappear,
whether or not any of them have been implemented by the appropriate
morphogenetic stimulus. Some competences exist in the adult tissues,
if we may so term the states of reactivity to hormone action, though these
effects are generally reversible."

It will be observed that these definitions all describe *effects*, and give
no hint as to the *mechanisms* whereby the effects are brought about.
Although the isolation, identification, and synthesis of biologically-active
chemical substances (e.g., vitamins, hormones), as well as the demonstra-
tion of the surprising effects produced (biotin is effective in a dilution of
1 part in 400-billion) fill us with wonder and admiration, and many scien-
tists have received or deserve to receive Nobel prizes for their work in
this and allied fields, all will end in verbiage unless we utilize these
splendid results to carry through and explain the mechanisms whereby
"organizers" organize, "evocators" evoke, and "inductors" induct.* It is
submitted that the concept of catalysts, and especially neocatalysts, sup-
plies the basis for the main mechanisms, although many other factors
are also simultaneously active and important.

To epitomize: a reasonable deduction from evidence in many fields is
that organizers, evocators, inductors, etc., operate largely by furnishing
specific particulate units which help to establish neocatalysts, by serving

* Many who know Goethe as a great poet and philosopher, are unacquainted with
the fact that in the Weimar edition of his works (126 volumes) his "Wissenschaftlische
Werke" occupy 12 volumes, and that in 1817 he introduced into biology the term
"morphology." His interest in science was profound, and the writer, therefore, would
like to present a rather free translation of the celebrated lines from "Faust" (Part I,
lines 1629 to 1646), where Mephistopheles, disguised as Dr. Faustus, is cleverly mis-
leading a student who, burning with idealism and ambition, has come to ask advice
from the great professor.

Mephistopheles:
              I would not guide you wrongly here.
              Considering this science rightly,
              'Tis hard to avoid the wrong road's terror;
              Much hidden poison, seeming sightly,
              Is used as medicine, in error.
              'Tis best to hear one single master,
              Swear by his words and dodge disaster.
              In general, stick to words and see
              How quick this gets you your degree
              From university or college.
Student:
              But words must hold some concept, knowledge.
Mephistopheles:
              Of course! Be not disturbed by that;
              For where ideas are lacking, flat,
              A word provides your rescue, pat.
              With words you'll argue full and fair,
              With words a system you'll prepare,
              In words you'll faithfully believe,
              Of every word each sliver you'll retrieve.

either as prosthetic groups or carriers, or as modifiers of existing catalysts. In some cases, the specific substances may alter the permeability of septa, thus affecting diffusion to and from catalysts; or they may serve as deflocculators of aggregates or as aggregators; or as inhibitors or removers of inhibitors by a detergent-like action; or they may preempt prosthetic groups or carriers and thus interfere with the formation of certain catalyst enzymes. For example, insulin inhibits a certain pituitary hormone. With the establishment of a nervous system, nerve influences are added to those of any circulatory system as an additional remote control. Nature explores and utilizes, where feasible, most—if not all—possible mechanisms, providing they lead to the formation and operation of catalysts whose chemical outputs yield a morphology and physiology advantageous to the organism.

From this point of view, *competence,* "the state of reactivity," simply means that the organism or part involved maintains or receives in operative state, and at the critical time, the specific molecules or other units which can serve as prosthetic groups, carriers, modifiers, etc., which enable it "to react to a given morphogenetic stimulus" (i.e., to specific molecules or other particles from organizers, evocators, inductors, etc.) "by determination and differentiation." Expressed in material terms, this means that specific molecules, etc., from cells, tissues, etc., interact with specific molecules, etc., from organizers, etc., to form neocatalysts which give new and often surprising quirks to the course of life processes. As new catalysts are formed modified, inhibited, released, or destroyed, new catalyst formers or modifiers may come into being and existing ones may vanish. The numerous, complicated, and mostly unknown chemical changes are understandable in material terms, at least in principle, at the catalyst level. The ability to form catalysts which lead to the clinical syndrome of cancer, seems in some cases to represent a competence which persists or is established or reestablished on ageing. In some cases, the cooperation of an invader (e.g., Rous virus) is essential or helpful.

Evolution

Evolution stems from the heritable transmission of advantageous neocatalysts and/or substances or units which determine the formation of advantageous neocatalysts. While most abnormalities are harmful, it occasionally happens that some deviation from the normal is advantageous under existing conditions. If such an advantageous abnormality is carried over to descendants, these may coexist with, or even replace the normal form, thus registering a step in evolution. If this happens in a state of nature, it is "natural selection" (Darwin) or the "survival of the fittest" (Spencer). Artificial selection or "breeding" has been practiced by men

from time immemorial by selecting plants or animals having desirable features, feeding them well, protecting them against competitors and predators, and controlling their matings. Such pampered individuals generally succumb in a state of nature—weeds quickly run out most cultivated plants. Surveying the passing landscape from a train, we see fields where plants and animals show the effects of human selection and protection; but the uncultivated areas show a highly-variable succession of wild grasses, herbs, shrubs, and trees, each group fighting to maintain its suitable habitat against competitors striving to take it.

Evolutionary competition may be observed at a variety of structural levels; e.g., reproductive (sterility vs. fecundity); sexual (success with the opposite sex); ability to meet milieu conditions (to become a predator or to foil predators, to secure food, to survive the vicissitudes of the environment). But none of the advantages any individual may develop can count in evolution unless it is passed on to descendants, and this transmission is accomplished mainly through the germ plasm (although with the higher animals the young are taught much by their parents). How are differences established in the germ plasm, and how are they transmitted to offspring?

The germ plasm consists of two portions: (1) the gene strings or chromosomes; and (2) the cytoplasm of the gametes. Geneticists have shown that individual genes may undergo "point mutations" whereby their influence on heredity is altered; and also that whole blocks of genes may be involved in heritable changes by inversions, deletions, translocations, and by the duplication of whole chromosomes (diploidy, polyploidy). On the basis of these proofs, and the inability to demonstrate satisfactorily that anything besides the gene strings dominates heredity, many geneticists were led to the view that cytoplasmic inheritance factors are nonexistent or of minor importance. However, it is now well-recognized that the cytoplasm contains not only many specific enzymes and molecules which play an important role in heredity, but also more complicated structures, e.g., plastids and mitochondria, which seem to be able to reproduce themselves. Recent work has shown that the cytoplasm may also contain a variety of particulate units which function somewhat as though they are free-living genes. These have been given such names as cytogenes, genoids, pangenes, cytoplasmic factors, plastogenes, neurogenes, plasmagenes—all of which show the reluctance of biologists to envisage the catalytic function of these units. In his paper "Beyond the Gene," Sonneborn [17] distinguishes three types of plasmagenes: (1) initiated by genes; (2) maintained but not initiated by genes; and (3) independent of genes. "Plasmagenes differ from nuclear genes, so far as now known, only in their location in the cytoplasm instead of the nucleus. To embrace the whole of what is now known concerning the determination of heredity traits, the theory

of the gene needs only to drop the specification as to the location of the gene. Stripped in this way, the theory of the gene reduces to the brief statement, 'Heredity traits are determined by self-duplicating, mutable units.' "

We are here maintaining the now old, but insufficiently appreciated view, that all these units (genes, plasmagenes, or whatever the words used to describe them) function either as biocatalysts to direct the chemical changes underlying all life processes, or else to produce essential units which form part of other biocatalysts with like function. The three types of plasmagenes mentioned above may be explained thus (using the same numerical order): (1) Genes may form, catalytically, duplicates of the whole or of any part of their specific surfaces. Two or more of these particles, if set free independently, may unite to form a specific catalyst unit. (2) Genes may catalyze the formation of particulate units which constitute essential parts (e.g., carriers, prosthetic groups) of enzymes or of other catalyst units. (3) Nongenic catalysts in the cytoplasm (e.g., mitochondria, plastids, enzymes) may catalyze the formation of subunits which can unite to make more of the nongenic catalysts, without the necessity of using any substances catalyzed by genes. These suggested mechanisms are by no means final, for the chemical changes catalyzed in a single cell are of formidable complexity, and the course of nature is not limited by our present inability to understand or to analyze every detail.

It is evident that we have in neocatalyst formation a mechanism whereby nongenic changes in bionts *might be transmitted to offspring*. Fortunately, this is *not* the usual and dominating procedure; otherwise we would have a veritable Babel of bionts. However, it does indicate how the inheritance of "acquired characteristics" *might* be understood, *in such cases where valid proof of such inheritance is found*. We may here repeat the wise statement of Conklin: [18] "The classic argument of the Weismannians was that *we can conceive of no mechanism* by means of which somatic changes can be carried back into the germ cells, *and therefore there is no such mechanism*. Now the fallacy of this argument is obvious; even if we could conceive of no suitable mechanism for this purpose, this does not preclude the existence of such a mechanism." Cunningham suggested [19] that "hormones" (used in Starling's original sense, "to stir up, or excite") could produce modifications in the soma by external stimuli, which "could affect the determinants in the gametes in such a way that the modifications would be inherited . . . real substances of the nature of special chemical compounds take the place of the imaginery gemmules of Darwin's theory of pangenesis or the 'constitutional units' of Spencer . . . there are two kinds of variation in evolution, one somatogenic and due to external stimuli, acting either directly on passive tissues or indirectly

through function, and the other gametogenic and due to changes in the chromosomes of the gametes which are spontaneous and not in any way due to modifications of the soma. Adaptations are due to somatogenic modifications, non-adaptive diagnostic characters to gametogenic muta- tions. It is a mistake to attempt to explain all the results of evolution by a single principle." *

The paleontologist, Osborn, stated in the third of a series of addresses on the origin of species, that the theories of Buffon (direct action of en- vironment), Lamarck, and Darwin, usually regarded as contradictory, are really complementary, for they all turn on the question of inheritance or transmission of individual adaptation, which may furnish the key to evolution. "On the affirmative side paleontology proves in the long run of geologic or secular time that both Buffon and Lamarck, as well as Darwin, were right in their main conceptions: organs which do not pay their way or are starved by disuse slowly drop out of the germ-plasm; vitally essential organs are either absolutely stable or progressive. Why not therefore concede the truth of the great conceptions of Buffon and Lamarck, even if immediate inheritance by the germ is disproved in the great majority of cases? Why not concede the still greater conception of Darwin, misled as he was to time by the marvelously rapid evolution of the germ-plasm witnessed in artificial selections?"

Disease and Drugs

Since normal physiology is mediated by normal catalyst systems func- tioning under the normal range of conditions, it is evident that any change in the catalysts or in their operations may result in deviations in physiology, which, when sufficiently marked to elicit a clinical syn- drome (concurrent aggregate of symptoms), is called disease. Many dis- eases are now known to involve catalyst-enzyme deficiencies or excesses, but we still await further analysis of disease processes at the catalyst level which dominates the chemical, and therefore the physiological, output. The ancient Egyptians had medical men who specialized in diseases of the heart, lungs, stomach, eyes, etc., and a survey of modern medical books indicates that diseases are still classified on the basis of the part of the body most noticeably affected and often bear the name of the medical man who first described the disease or its cause to his fellows (Bright's disease, Hansen's disease). Many of the correct observations of earlier times are being more clearly understood in the light of present knowledge. Galen recognized four *temperaments:* (1) sanguine; (2) phleg- matic; (3) bilious or choleric; (4) melancholic. Later medical men spoke

* A review of Cunningham's book, by M. F. Guyer, appeared in *The Journal of Heredity,* 14, 136-8 (1923).

of *diathesis,* a predisposition to certain forms of disease. Results were described, not causes or mechanisms. A hemorrhagic diathesis was defined as "a morbid condition marked by lessened coagulability of the blood and an abnormal liability to bleed at slight wounds." Modern genetics has shown that this "diathesis" is carried by heredity, but does not indicate the physicochemical mechanisms responsible for the symptoms.

These older clinical aspects point up the fact that patients differ widely in potentialities, so that a particular patient must be treated, not a stereotyped "disease." There are no rigid rules telling us how any individual will respond to an infection or a drug, even though experience has taught us what is usual. These mainly genetic differences in individuals appear in colonies of animals kept for biological experiments and are found even in litter mates, so that they are a factor in judging the results of animal experiments. At a conference on "Animal Colony Maintenance" (New York Academy of Sciences, 1945), Dr. Edmund J. Farris (Wistar Institute) stated that in his undergraduate days "type" experiments in mammalian physiology "hardly ever agreed with the textbook picture. This was common classroom experience." Many scientists still do not realize that nature, especially in highly complex biological systems, refuses to follow textbook rules rigidly.

It is only comparatively recently that discoveries in various sciences have enabled us to trace the causes and processes of disease to lower and lower structural levels and to understand that perversions of biocatalysts and their operations may often confront the practicing physician with a puzzling clinical picture. Less than 100 years ago (1857), Pasteur read his epoch-making papers on lactic acid and alcoholic fermentation and told the French Emperor (1863) that his ambition was "to arrive at the knowledge of the causes of putrid and contagious diseases." When, in 1876, Robert Koch confirmed the earlier (1863) observation of Devaine that minute rods (Bacteridia) exist in the blood of animals dead of anthrax and are causative agents of this disease, Pasteur confirmed Koch's work and convinced those who opposed it. Doubters were confounded when, in 1881, in the farmyard at Melun, twenty-five sheep protected by Pasteur's vaccine survived massive inoculation with anthrax, while every one of the twenty-five unprotected control animals perished.

Bacteriology and parasitology have done much in identifying and studying microorganisms involved in diseases, and often in isolating chemical substances from which successful vaccines may be produced. Biochemists have identified, synthesized, and studied the effects of numerous chemical compounds potent biologically, many of them even in mere traces. But an adequate consideration of disease from the biocatalyst level is still to be developed. Gout, of old "uric acid diathesis,"

was defined as "a condition of the system in which uric acid is deposited in the joints." Osler, in his great book,[20] goes far beyond this banal statement, and refers to the catalytic aspect. "Uric acid, in the body almost completely in the form of urates, is the end-product of purine metabolism in man, just as urea is the end-product of nitrogenous substances of amino-acid and pyramidine origin. In man the end-product of purine metabolism is excreted as urate, there being no enzyme in man to break it down further to allantoin, as occurs in most mammals." For most diseases the following information is given: etiology; pathology (where known); symptoms; prognosis; diagnosis; treatment. With this vast amount of practical knowledge there is included for some diseases, prophylaxis and epidemiology, but only rarely pathological physiology, although from the latter there emerges at higher structural levels the syndrome of the clinician. Progress in medicine demands increasing knowledge of *mechanisms* as well as of *results*. Biocatalysts dominate biochemical change, the basis of physiology.

Some knowledge of the catalytic differences between animals is given by studies of detoxication, the mechanism whereby the organism eliminates certain poisonous substances. For example, fowls elimniate phenylacetic acid and its homolog, benzoic acid, by combining them with ornithine, whereas all mammals combine benzoic acid with glycine and eliminate it as hippuric acid. Sherwin and Thierfelder[21] found that though the lower animals, including monkeys, combine phenylacetic acid with glycine and eliminate it as phenaceturic acid, man couples phenylacetic acid with glutamine and eliminates it as phenylacetylglutamine. Power (Fordham University), experimenting with a chimpanzee, found[22] that this "man-ape" eliminates phenylacetic acid by the same biochemical mechanism as man. This result indicates a definite relationship between chemical mechanisms and the position of the animal in the taxonomic scale, just as has been found in the biochemistry of muscle.

Most of the apparently conflicting views regarding the protean group of diseases called "cancer" are resolved if we consider the catalytic level of biological happenings; for the cellular catalysts (genes, enzymes, symbionts) dominate the chemical and physicochemical changes which lead a cell to abnormal self-duplication *and* the invasion of surrounding tissues. Since cancer cells duplicate themselves as cancer cells, it is evident that a heritable change may arise from quite a number of different initial causes, e.g., burns (*kangri* cancer of Kashmir), radiation (x-rays and radium), specific chemicals (3,4-benzpyrene, methylcholanthrene), biological invaders (Rous "virus," liver fluke), the "milk factor" in mice (apparently a "virus").* The initial cause may be transmitted on cell

---

* See Editor's Note on p. 327 *et seq.*

division as a symbiont, or it may drop out after having established a neocatalyst pattern which is heritable, either by modifying a gene or establishing a self-duplicating unit within the cell. Irrespective of the precise mechanism involved, the result is called *cancer* if the cell goes on to abnormal duplication *and* the destructive invasion of surrounding tissue, often accompanied by the migration of cells from the cancer mass to initiate new growths elsewhere (metastases). Further aspects of this subject are treated by the writer in "Life—Its Nature and Origin." [23] In some respects, plant galls and fasciation in plants (see e.g., paper by White [24]) resemble cancers.

Drugs. From the most ancient times people of all races and nations have acquired a great fund of knowledge about drugs and remedies, even though their ignorance of the operative mechanisms often misled them into doing unnecessary things and making incantations. The Chinese Emperor Shên Nung (2838 to 2798 B.C.) is said to have written the first *Pên T'sao* ("Native Herbs") in three volumes which describe 365 drugs. In the reign of Djoser (Zoser), first Pharaoh of the Third Dynasty (Old Kingdom), who ruled from 2780 to 2760 B.C., the great engineer and architect, Imhôtep, did much to advance the healing art, and was deified as the God of Medicine, a name later given by the Greeks to Aesculapius. Our modern pharmacopeia is indebted to many sources: to China for stramonium, chaulmoogra oil, and ephedrine (*ma huang*); to the American Indians for cascara sagrada (from the buckthorn); to South American natives for quinine. The African pygmies are said to pursue a huge elephant for days, peppering his hide with arrows dipped in strophanthus until this potent heart poison brings him down. In more recent time, the active chemical substances have been isolated from many drug plants—morphine, strychnine, atropine. In 1868 Brown and Fraser found that by introducing a methyl group into strychnine, theobaine, or brucine, the tetanizing action of these drugs was changed into a paralyzing one (Sir T. C. Allbutt). Later, synthetic drugs appeared: antipyrine (1883), phenacetin (1887), aspirin 1899), and salvarsan (1912). The medicinal value of the long-known sulfonamide was discovered in 1939, and still more recently penicillin, streptomycin, and many other potent products have been found and successfully utilized.

Many drugs, especially those which act in minute amounts, operate by affecting specific enzyme systems. The ideal antibiotic is one which will inactivate or destroy the essential catalysts of an invading pathogenic organism without too much damage to the catalysts of the patient. The ratio between the minimum curative dose and the minimum lethal or poisonous dose is known as the therapeutic index, and in safe drugs the spread is a wide one. Organisms often react to the presence of drugs by developing immunity to them, so that the initial dose should be as potent

as conditions permit. In some cases, bacteria assume defensive forms, having a changed morphology and behavior, but still capable of producing disease. The extreme complexity of organisms means that drugs may produce effects far removed from their initial point of action. Thus, inhibition of the catalysts of the respiratory center by cyanide or hydrogen sulfide brings quick death.

REFERENCES

1. Downs, C. R., U.S. Patent 1,604,739 (Oct. 26, 1926).
2. Pliny, *Historia Naturalis*, xxiv. Chap xiv, sec. 42.
3. Brody, S., "Bioenergetics and Growth," p. 111, New York, Reinhold Publishing Corp., 1945.
4. Bergmann, M., and Fruton, J. S., *Ann. N. Y. Acad. Sci.*, 45, 409-23 (1944).
5. Rahn, O., *Growth*, 4, 77 (1940).
6. Alexander, J., and Bridges, C. B., in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. II, p. 13, New York, Reinhold Publishing Corp., 1928.
7. Breinl, F., and Haurowitz, F., *Z. physiol. Chem.*, 192, 45 (1930).
8. Alexander, J., *Protoplasma*, 14, 296 (1931).
9. Mudd, S., *J. Immunol.*, 23, 423 (1932).
10. Hooker, S. B., and Boyd, W. C., *J. Immunol.*, 21, 113 (1931).
11. Miall, S., ed., *Chemistry & Industry*, editorial (Sept. 30, 1932).
12. Alexander, J., "Life—Its Nature and Origin," p. 89, New York, Reinhold Publishing Corp., 1948.
13. Sevag, M. G., "Immuno-catalysis," Springfield, Ill., Charles C. Thomas, Pub., 1945.
14. Needham, J., "Biochemistry and Morphogenesis," p. 686, Cambridge, Engl., Cambridge Univ. Press, 1942.
15. *Ibid.*, p. 685.
16. *Ibid.*, p. 682.
17. Sonneborn, T. M., "Beyond the Gene," *Am. Scientist* (Jan., 1949).
18. Conklin, E. G., "Heredity and Environment," 3rd ed., Princeton, N. J., Princeton Univ. Press, 1919.
19. Cunningham, J. T., "Hormones and Heredity," London, Constable and Co. Ltd., 1921.
20. Osler, W., "The Principles and Practice of Medicine," 15th ed., 1944.
21. Sherwin and Thierfelder, *Ber.*, 47, 2630 (1914).
22. Power, F., *Proc. Soc. Exptl. Biol. Med.*, 33, 598 (1936).
23. Alexander, J., "Life—Its Nature and Origin," Chap. 10, New York, Reinhold Publishing Corp., 1948.
24. White, O. E., *J. Heredity*, 36, 11-22 (1945).

ADAPTIVE FORMATION OF ENZYMES
AND INFLUENCE OF NITROGEN NUTRITION
ON THE ENZYMATIC ACTIVITY OF CELLS

Artturi I. Virtanen

Biochemical Institute, Helsinki, Finland

ADAPTATION IS a phenomenon characteristic of all living beings. Without adaptation it would be hard to understand how living things of some sort exist under all possible conditions on the globe. The life processes of cells depend on enzymes which, according to our present knowledge, are proteins. Whatever other groups may be joined to protein in the different enzymes are beyond the scope of this paper. Since metabolism reactions are catalyzed by enzymes, in the study of the adaptation phenomenon it is of special interest to examine the changes taking place in the enzyme system of cells under diverse conditions.

Adaptive Enzyme Formation and Nutrition

Primarily, attention is fixed on the role of nutrition in the enzyme system of cells. Experiments in this direction have been carried out in the laboratory of the Biochemical Institute since the 1920's. In his doctoral thesis Karström[1] classified enzymes into two groups, depending on how their formation is related to nutrition. The *constitutive* enzymes are formed independently of the composition of the nutrition of the cell, the *adaptive* enzymes only when the substrate for the particular enzyme provides the nutrition of the cells. Thus, for instance, many bacteria produce lactase, maltase, and sáccharase only if lactose, maltose, and saccharose, respectively, are present in the nutrient solution. When such cells are transferred to nutrient solutions lacking the said disaccharides the respective enzymes disappear upon division of the cells.

This remarkable dependence of enzyme formation on nutrition is highly interesting both chemically and biologically. Without discussing in detail the various aspects opened up by the concept of adaptive

255

enzyme formation we here report on two important observations on the formation of adaptive enzymes, made in the Institute laboratory during the 1940's.

Firstly, Virtanen and Winkler [2] showed that the formation of lactase in *E. coli* in lactose nutrient solution is neither a mutation phenomenon nor a selection, but an occurrence affecting all cells. This evidence was furnished in the following way: A coli-strain which was repeatedly cultivated in saccharose nutrient solution (90 times during 180 days) and which did not show any lactase activity, was transferred to liquid lactose-agar nutrient (45°C). The lactose-agar contained sufficient bromcresol purple indicator to color the solution deep violet. Different dilutions of bacteria were made and the contents of the tubes were poured into Petri dishes. After two days at room temperature, colonies were formed which were encircled by a yellow zone because of acid formation from lactose (bromcresol purple changes to yellow at pH less than 6.2). In a parallel experiment with saccharose-agar, approximately as many colonies were formed, and the color of the indicator had also changed to yellow around all colonies. Table 1 records one of the experimental series. This result

TABLE 1. NUMBER OF COLONIES (PER SQ. CM) OF LACTASE-FREE *E. coli* (K₃) ON
LACTOSE AND GLUCOSE-AGAR

Dilutions

		II	III	IV	V
Lactose-agar	Countless numbers	257	74	40	2.1
Glucose-agar	Countless numbers	281	92	44	2.6

is difficult to interpret except by assuming that practically all the cells have produced lactase on lactose-agar.

Secondly, the same workers made the noteworthy observation that lactase formed through adaptation of short duration disappears rapidly after the lactase-containing cells are transferred to lactose-free nutrient solution; but not immediately after an adaptation of long duration.[2] When *E. coli* was cultivated two hundred times successively in lactose nutrient solution and allowed to grow for two days each time—the number of generations was then already in the thousands—lactase did not disappear until after the ninth transfer to lactose-free nutrient solution (Table 2).

The above results seem to suggest that through adaptation of long duration the enzyme becomes more and more stable. As both these and Spiegelman's [3] experiments prove that the proteins of the enzymes may interchange in adaptive enzyme formation, the results could be interpreted to mean that through the effect of a long-standing adaptation the

enzyme protein becomes more and more constant. Mutation of individual cells is, however, possible. All explanations of this phenomenon are, however, speculative so far. In this connection, reference is made to the lecture given by the writer in 1947, discussing this complex problem.[2]

The starting point for our latest research on enzyme formation and on the dependence of enzyme activity of the cells on their protein content, was the idea advanced by Virtanen in 1942 [4] that the entire protein of young, active cells is practically enzyme protein or "living" protein.

TABLE 2. STABILITY OF LACTASE IN *E. coli* DEPENDING ON DURATION OF ADAPTATION

(*E. coli* Transferred from Glucose to Glucose Nutrient Solution for 1 to 9 Times After Training in Lactose Nutrient Solution for 1 to 199 Times)

Number of Transfers in Lactose Nutrient Solution	Number of Transfers in Glucose Nutrient Solution						
	1 Lf 10^4	2 Lf 10^4	3 Lf 10^4	5 Lf 10^4	7 Lf 10^5	8 Lf 10^4	9 Lf 10^4
1	0						
10	0						
20	21	0					
30	72	0					
40	72	10					
161	170?		69	68	44		0
181	78	66	70	71	14	0	
199	105	62	37	44	24	33	0

Lf 10^4 = about 700 in lactose nutrient solution.

The protein content of cells would thus control the enzymatic activity of cells. As the protein content is a limiting factor in enzyme formation, adaptive enzyme formation means a most effective utilization of proteins. Thereby the cell is able to form, when forced to do so, new enzymes at the expense of some other protein which is not indispensable at the moment. The greater the ability of an organism to form various enzymes by adaptation, the better its chances of survival in different nutritional conditions and in competition with other species. Adaptable organisms are not the first to disappear from the globe.

Nitrogen Nutrition and Enzyme Activity

If the above concept that practically all the cell protein belongs to enzymes holds true, it is to be expected that as the protein content falls, all enzymes may have lessened activity, or only some may retain their activity, while others may lose most or even all of it. This question can be approached experimentally because the protein content of unicellular forms can easily be reduced by cultivating the organisms in low-nitrogen solutions (bacteria) or by suspending the cell mass with normal protein
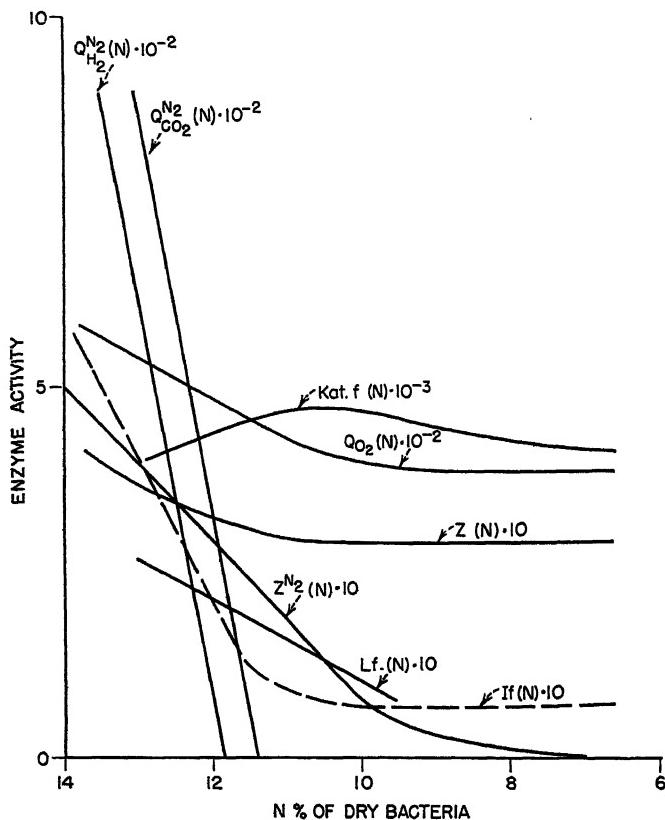
Figure 1. Changes in the activity of different enzymes as affected by the N-content of *E. coli*. All activities calculated on the basis of N-content of cells.

Kat. f (N)·10^{-3} — catalase effect (Virtanen and De Ley).
If (N)·10   — saccharase effect (Virtanen and De Ley).
Q_{O_2} (N)·10^{-2} — respiration (De Ley).
Z (N)·10   — aerobic acid formation (De Ley).
$Z^{N_2}(N)$·10   — anaerobic acid formation (De Ley).
$Q_{H_2}^{N_2}(N)$·10^{-2} — H_2 formation from formic acid (De Ley).
$Q_{CO_2}^{N_2}(N)$·10^{-2} — CO_2 formation from formic acid (De Ley).
Lf (N)·10   — lactase effect (Virtanen and Winkler).

content in sugar solution under vigorous aeration (yeasts). In this manner Virtanen and De Ley [5] have obtained from *E. coli* cell masses with very different N-content, the extreme values being about 13 and 6.5 per cent N. The catalase activity of the cells proved to be almost independent of their N-content, but the saccharase activity fell abruptly while the N-content lowered even slightly. When the N-content of the cells decreases

from 13 to 11 per cent, there is only about 10 per cent left of the saccharase activity. This corresponds approximately to the constitutive saccharase in *E. coli.* Saccharase is no longer formed by adaptation in low-nitrogen cells. These observations imply that enzymes which are indispensable to life in all conditions and which at the same time are constitutive, are retained rather well while the protein content of the cells decreases. On the other hand, enzymes which are not indispensable to life in all conditions and which are active only when a specific sub-
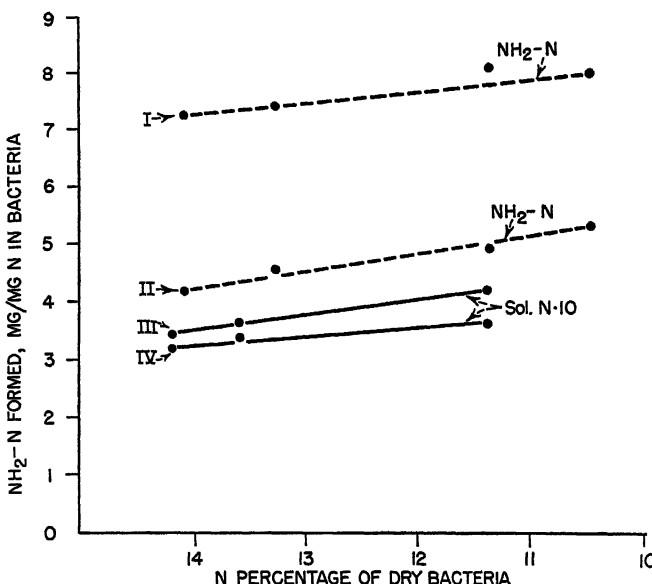
Figure 2. Formation of soluble- and amino-nitrogen from casein by *E. coli.*

Experiment 1 { Curve I: NH₃–N formed during 20 days.
Curve II: NH₃–N formed during 5 days.
Experiment 2 { Curve III: Soluble N formed during 17 days.
Curve IV: Soluble N formed during 6 days.

strate provides the nutrition of the cell, lose correspondingly the major part of their activity or perhaps disappear entirely with decreased cell protein. These enzymes are often adaptive in microorganisms. The idea advanced by Virtanen as to the enzymic character of the cell proteins is supported by these experiments.

Similar experiments have also been made with different microorganisms in regard to many enzymes and enzyme systems. In his doctoral thesis De Ley [6] reports on noteworthy findings in respect to respiration and anaerobic fermentation with the same strain of *E. coli* used earlier by Virtanen and De Ley.[5] Respiration weakened relatively little when

the N-content of the cells dropped to half. Anaerobic fermentation surprisingly came to an end when the N-content fell from 13 to 8 per cent. The formic hydrogenlyase disappeared entirely when the N-content fell only to 11 per cent. Figure 1 presents these results.

Virtanen and Winkler [7] have observed that the proteolytic enzyme system of *E. coli* is retained very well when the N-content of the cells is lowered (Figure 2).

Virtanen and Kokkola have studied the influence of nitrogen nutrition on the content of protease excreted into the nutrient solution by *Pseudomonas fluorescens*. This enzyme or enzyme complex maintains its activity very well when growing in nutrient solutions of very different ammonium
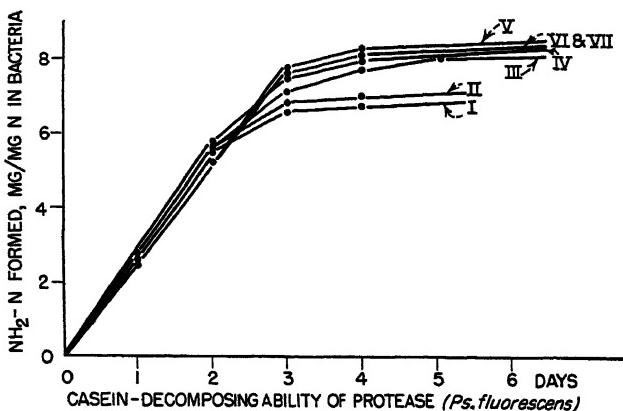
Figure 3. Dependence of the casein-decomposing ability of protease excreted by *Pseudomonas fluorescens*, on the ammonium sulfate concentration of the nutrient solution.

	Curves						
	I	II	III	IV	V	VI	VII
(NH₄)₂SO₄	0.05	0.1	0.5	5.0	10.0	15.0	25.0
(gm/5-1 nutr. soln.)							

sulfate concentrations (Figure 3). Respiration of this organism is also well retained while the N-content of the cells decreases from 13.7 to 5.4 per cent. Anaerobic fermentation, on the other hand, decreases noticeably.

Experiments using *Torula utilis*-yeast were carried out with Aejmelaeus. These revealed that while the N-content of cells decreased to about half, or from nearly 10 to slightly below 5 per cent, the respiration intensity and catalase activity were very well retained. A small rise could be noted even when the nitrogen content decreased. Saccharase, too, which is a constitutive enzyme in *Torula*, retains its high activity surprisingly well (Figure 4). Only when the N-content is lowered to below 4 per cent—which can be accomplished by adding β-alanine to strongly aerated

yeast suspension in sugar solution—are the activities of all these enzymes abruptly decreased. The viability of cells is by then very feeble.

Weakening of the anaerobic fermentation in *Torula* when the N-content of the cells decreases to below 8 per cent is significant. The observations made both with facultatively anaerobic bacteria and *Torula*
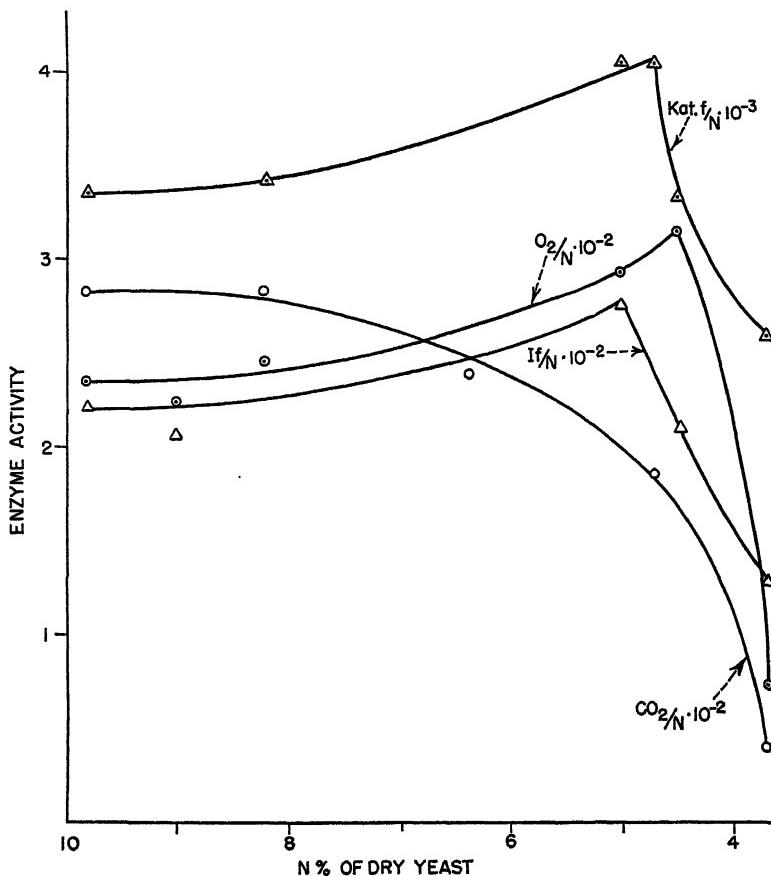
Figure 4. The respiration and fermentation intensities and catalase and saccharase activities of *Torula* yeast.

yeast on the weakening of anaerobic fermentation, suggest that the acquisition of energy by fermentation is difficult for these organisms and that much nitrogen nutrition is required for the process.

The kind of changes in the amino acid composition of cell proteins brought about by a very low nitrogen concentration of the nutrient solution has not yet been ascertained. But in experiments with *Ps. fluorescens* both the nitrogen fractions soluble and insoluble in trichloracetic acid

were determined in ammonium sulfate concentrations varying from 25 to 0.05 gram per 5-liter nutrient solution. The results illustrated by graphs in Figure 5 show that in the ammonium-ion concentration optimal to growth (10 grams $(NH_4)_2SO_4$ per 5 liters) about 85 per cent of the total nitrogen content of the cell was insoluble in trichloroacetic acid. Proteins belong to this fraction. As the ammonium sulfate concentration in the nutrient solution decreases and the N-content of the cells falls, the percentage of total nitrogen of the fraction insoluble in trichloroacetic acid diminishes. The differences, however, are not very great but still noticeable. As the fraction soluble in trichloroacetic acid does not contain
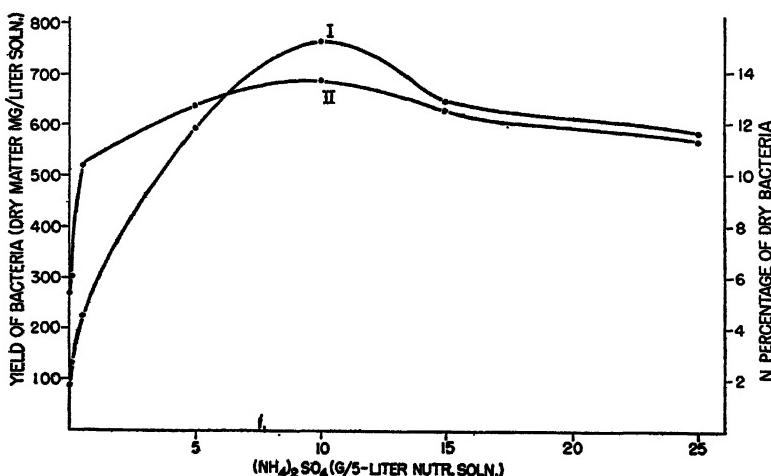
Figure 5. Dependence of the growth and N-content of *Ps. fluorescens* on the ammonium-sulfate concentration of the nutrient solution.
Curve I: Yield of bacteria (dry matter) per liter of nutrient solution.
Curve II: N-content of bacteria (% of dry matter).

proteins, it would be more correct to express the enzyme activity as a function of the nitrogen insoluble in trichloroacetic acid, rather than of the total nitrogen. We have, however, maintained the latter expression since the nitrogen fraction insoluble in trichloroacetic acid was determined in one experiment only. In any case, it is interesting that the amount of nitrogen nutrition, in addition to the nitrogen content of the cells, affects the mutual relations of different nitrogen compounds. It is also noteworthy that the superoptimal ammonium-ion concentration (25 grams $(NH_4)_2SO_4$ per 5 liters) which begins to retard the growth of cells, also lowers the proportion of the nitrogen fraction insoluble in trichloroacetic acid (in per cent of total N). (See Table 3.)

All the data recorded above show that by means of nitrogen nutrition it is possible to affect strongly the enzymatic machinery of cells. This fact

TABLE 3. EXPERIMENT WITH *Pseudomonas Fluorescens*

(Percentage of N-Fraction, Insoluble in Trichloroacetic Acid, in Total N of Cells, and the Dependence of This Fraction on the Ammonium-Sulfate Concentration of the Nutrient Solution)

Nutrient Solution $(NH_4)_2SO_4$ g/5-1	N-Content of Bacteria, % of Dry Matter	N-Fraction Insoluble in Trichloroacetic Acid, % of Total N in Bacteria
25	11.5	76
15	11.8	81.8
10	12.4	85.9
5	11.9	84.3
0.5	10.1	77.2
0.1	6.1	76.4
0.05	5.4	75

opens new vistas in various directions. The amount of nitrogen nutrition may induce formation of new species either by mutation or slow evolution. Observations on microorganisms may throw new light on the significance of low-nitrogen nutrition in human and animal feeding, in which enzyme systems may be similarly affected; it is difficult to demonstrate these effects in higher organisms.

REFERENCES

1. Karström, H., *Ann. Acad. Sci. Fennicae,* **33A** (2), (1930).
2. Virtanen, A. I., *Svensk Kem. Tid.,* **60,** 23 (1948).
3. Spiegelman, S., *Cold Spring Harbor Symposia Quant. Biol.,* **11,** 256 (1947).
4. Virtanen, A. I., *Suomen Kemistilehti,* **15B,** 22 (1942).
5. ———, and De Ley, J., *Arch. Biochem.,* **16,** 169 (1948).
6. De Ley, J., "Over de fermenten van stickstof-arme Bacterium coli," Thesis for doctorate, Gent, 1949.
7. Virtanen, A. I., and Winkler, U., *Acta Chem. Scand.,* **3,** 272 (1949).

SOME ASPECTS OF THE INTERACTION
OF DRUGS AND ENZYMES

Frederick Bernheim

Professor of Pharmacology, Duke University Medical School, Durham, N. C.

IF A PHARMACOLOGIST is defined as a man interested in the mode of action of drugs, he should be a perfect amalgam of a physical, colloidal, organic and biochemist—a physiologist, bacteriologist and probably a histologist as well. Few, if any, such individuals exist, and as a consequence, perhaps, there is no complete explanation of the action of any drug. Moreover, it is possible that no complete explanation will be forthcoming until all the factors involved in the life of the cell and the whole organism are known. All that the pharmacologist can do at the present time is to correlate data obtained by a number of individuals using different techniques, in order to obtain an approximation of drug action. The first necessity is, of course, a precise description of the effects produced in animals by the drug, with particular attention to the exact locus of action. Once such information is available, it is sometimes possible to isolate the groups of cells primarily affected and to study the action of the drug in a more quantitative way. Results from such experiments may indicate whether the drug is altering the physical or colloidal state of the cell membrane, or is penetrating into the cell to affect one or more of the numerous catalytic activities which are the basis of the cell's metabolism. If the latter alternative seems more likely, various recognized biochemical techniques can be used in order to determine which of the catalysts are acted upon by the drug.

It should be noticed that this approach is essentially analytical in the sense that no generalization or hypothesis is involved. All the older drugs were discovered empirically, and the pharmacologist has a tremendous backlog of work to explain their action. It is, however, unfortunately true that many of the newer drugs have also been discovered in the same way, and therefore pharmacology must still be considered incomplete. The ability to predict effects, as well as to analyze them, is the most

important characteristic of a scientific discipline. Thus, besides trying to find an explanation for the bacteriostatic action of the sulfa drugs, for example, the pharmacologist, knowing the metabolic and growth characteristics of an organism, should be able to indicate the compound which would specifically inhibit it. Predictions based on this type of reasoning have met with some success. Metabolic antagonists and vitamers of various types inhibit growth of bacteria, and in some instances of animals as well, but no very potent therapeutic agent has as yet been developed. Predictions based on purely chemical reasoning have met with more success, as is evident in the development of the newer analgesics.

Most drugs are changed to some extent in the body, and it is possible to look for the enzymes which catalyze the change. This search often leads to interesting but occasionally misleading results. The interest lies in finding an enzyme which attacks a drug, with a chemical configuration quite different from any metabolite hitherto found in the cell under normal conditions. This may lead to bemused speculation about the normal function of the enzyme, whether it has such, or whether its presence signifies some haphazard mutation, or a response to conditions, such as feeding habits in the past, which no longer prevail. Then, after all, the enzyme just discovered may actually play only a minor part in the overall metabolism of the drug in the body. This may be caused by permeability conditions which prevent the drug from coming in contact with the enzyme, or to physical conditions which concentrate the drug in cells which do not contain the enzyme.

Cholinesterase

It is the purpose of this discussion to illustrate some of the foregoing points by specific examples. The discovery by Loewi of the part played by acetylcholine in the transmission of nervous impulses, and of the inhibition of the hydrolysis of acetylcholine by eserine, was the initial impetus to the study of the effect of drugs on enzymes. It remains true today that the best correlation between the physiological action of a drug and its effect on an enzyme is found in the cholinergic system, particularly the autonomic part. The primary hypothesis is that any drug which inhibits the cholinesterase should give prolonged cholinergic effects when it is injected into the body. Certain modifying assumptions must, however, be made. It is known that acetylcholine is released at nerve endings and then combines momentarily with receptors in the muscle, after which it is hydrolyzed by the esterase. For the latter process to occur, acetylcholine must combine with the enzyme.* There are thus

* A discussion of recent work in this field by David Nachmannsohn is given in J. Alexander's "Life—Its Nature and Origin," New York, Reinhold Publishing Corp., 1948.

two entities in close geographical proximity, both capable of combining with acetylcholine and therefore having presumably similar chemical configurations. A drug which inhibits the cholinesterase *in vitro* will also have an affinity for the receptors in the muscle when it gains access to the myoneural junction. Whether it combines with esterase or receptor will depend on the relative affinity it has for the two loci. If it combines with the esterase, transmission will be facilitated; and if it combines with the receptor, transmission will be depressed. Consequently, a drug which inhibits the esterase *in vitro*, such as atropine, may actually prevent transmission.

Another factor which has to be considered is the innervation of the adrenal medulla. Adrenaline is released from this gland as a result of stimulation of a cholinergic nerve. Cholinesterase inhibitors may therefore cause an increased release of adrenaline into the blood stream. Adrenaline, in general, has effects opposite from those of acetylcholine, and therefore the overall picture of increased cholinergic activity in the autonomic system, expected after the injection of the inhibitor, is blurred by the simultaneous increase of adrenaline effects.

There are two enzymes which hydrolyze acetylcholine, the so-called pseudo- and true cholinesterases. They differ in distribution; the former is present in the plasma of most animals, the latter in the red cells and the brain. Other organs contain mixtures. The enzymes differ also in their substrate specificity, for the pseudo-esterase hydrolyzes benzoylcholine but the true esterase does not. The concentration of acetylcholine for optimal hydrolysis is lower for the true than for the pseudo-esterase. Any drug which inhibits one enzyme will also inhibit the other. It is conceivable that such a drug introduced into the body may be bound by the blood enzymes which apparently have no immediate critical physiological functions, and that little, or none, of the drug is available to inhibit the enzyme in the brain, myoneural junctions, or other critical sites. The drug, therefore, may not produce symptoms until the blood enzymes have been saturated. Another factor related to this is that many enzymes are present in the body in concentrations higher than necessary for normal function. This is apparently especially so in the case of the brain cholinesterase. Mazur and Bodansky[1] have shown that in rabbits this enzyme must be inhibited more than 40 per cent by diisopropylfluorophosphate before symptoms appear. Even an 80 per cent inhibition may be compatible with life. Finally, the same enzyme in different animals may show different sensitivities to the drug. Table 1, taken from Mazur and Bodansky, illustrates this point.

Despite these factors, the correlation between the *in vitro* and *in vivo* actions of diisopropylfluorophosphate is excellent. Not only do the different sensitivities of enzymes from different sources demonstrated *in vitro*

appear *in vivo* after the animal is exposed to the drug, but most of the symptoms expected from increased cholinergic effects, such as tremors, salivation, respiratory difficulty, and gastrointestinal spasm are seen. Moreover, there is excellent correlation between the esterase-inhibiting potency and the toxicity of the group of compounds related to diisopropyl-

TABLE 1. THE EFFECT OF DRUGS ON THE CHOLINESTERASE FROM DIFFERENT SOURCES

Drug	Animal	Neg. Log. Conc. to Produce 50% Inhibition		
		Serum	R.B.C.	Brain
Diisopropylfluorophosphate	Rabbit	4.1	5.2	5.5
	Monkey	7.8	5.5	5.5
	Man	7.7	5.4	6.0
Eserine	Rabbit	5.9		
	Man	6.7		

fluorophosphate, which strongly suggests that their entire effect in the body can be explained by the inhibition of the cholinesterase. The data in Table 2, taken from Jones *et al.*,[2] show this. If the figures for inhibition and toxicity are plotted, a straight line is obtained. With eserine or prostigmine the correlation is almost as good, with the exception of their action on certain spinal reflexes, which varies in different animals and which may find its explanation in the degree of esterase inhibition; for

TABLE 2. CORRELATION BETWEEN THE EFFECT OF DIFFERENT DRUGS ON THE CHOLINESTE-
RASE AND THEIR TOXICITY

Formula	Molar Conc. for 50% Inhibition of Rat-Brain Esterase	LD$_{50}$ for mice μM/Kg
$(C_2H_5O)_2POOPO(OC_2H_5)_2$	6.3×10^{-9}	4.0
$[(C_2H_5O)_2OPO]_2POCH_3$	1.0×10^{-8}	9.5
$FPO[OCH(CH_3)_2]_2$	4.3×10^{-7}	48.6
$ClPO[OCH(CH_3)_2]_2$	2.0×10^{-6}	146.0
p-$ClC_6H_4PO(OC_2H_5)_2$	1.6×10^{-5}	562.0

it is well known that if too much acetylcholine accumulates, depression of transmission occurs.

Many other drugs inhibit the cholinesterase but this can only be a partial explanation of their effects in the body. Morphine belongs in this category. It was first shown to inhibit the esterase by Bernheim and Bernheim,[3] and Eadie[4] demonstrated that it competed with acetylcholine for the enzyme. Many of the peripheral effects of the drug can be ex-

plained by this inhibition, but the drug also has central effects which
in some cases counteract the peripheral, so that a precise analysis is often
difficult. Its main therapeutic action, namely analgesia, cannot be ex-
plained directly by enzyme inhibition. It is interesting, however, that
two other powerful analgesics, meperidine and methadon, also inhibit
the esterase. Gross et al.[5] believe that by this mechanism the three drugs
increase the adrenaline output of the adrenal medulla and the vasocon-
striction thus produced is responsible, at least in part, for the analgesia.
They base their evidence on the fact that drugs are much less effective
in adrenalectomized animals and that drugs which antagonize cholinergic
effects decrease the analgesic action. Finally, it is well-known that mor-
phine and its derivatives are centrally-acting emetics. Kuhn and Surles[6]
have shown that apomorphine is a better inhibitor of the esterase than
morphine which, in turn, is better than dilaudid or codeine. This is the
order of their emetic effectiveness and suggests that inhibition of the
enzyme in the region of the vomiting center may be part of the mech-
anism by which emesis is initiated.

Cytochrome Oxidase

The cytochrome oxidase is probably more widely distributed in the
animal body than is the cholinesterase and it performs an equally vital
function, although in a different way. The cholinesterase is not essential
to the life of any cell, but only to the life of the organism as a whole.
Inhibition of it causes a disorganization which may result in death. The
cytochrome oxidase, on the other hand, is essential for the activity of
the individual cell, and death of the organism occurs when a drug
inhibits the enzyme and thus the oxidative metabolism of the cell.
Cyanide is an example of such a drug. It combines with the cytochrome
oxidase so that the latter can no longer oxidize reduced cytochrome C in
the presence of oxygen. This means that all the metabolites which are
oxidized through the cytochrome system remain reduced and the cell is
asphyxiated even though plenty of oxygen is available. The symptoms
of cyanide poisoning should be the same as those of oxygen lack, with
the exception that in the latter case the cell environment is depleted of
oxygen and the cytochrome C remains reduced for that reason. Death
from cyanide poisoning is more rapid than death from asphyxia because
the cells in the latter condition can utilize all the oxygen in the blood,
whereas in the former death occurs with completely oxygenated hemo-
globin. The symptoms of asphyxia and cyanide poisoning are otherwise
identical. They include respiratory stimulation through the carotid body,
convulsions, adrenaline release, rise in blood lactic acid and finally death
due to cardiac failure and paralysis of the medullary centers. It is thus

possible to conclude that the symptoms of cyanide poisoning are attributed entirely to the inhibition of one enzyme.

Sulfhydryl Enzymes

When drugs, such as cyanide, act primarily on one enzyme, analysis of their action in the body in terms of this effect can be fairly precise. But if a drug is known to inhibit a large number of enzymes, analysis becomes difficult; all one can really say is that the toxicity of the compound is thereby explained. Mercury, whether in the inorganic or organic form, is a drug of this kind. Barron and Singer [7] have shown that mercury combines with sulfhydryl groups on enzymes and that in low concentrations this is its only action. By studying the action of mercury, under standard conditions, on a large number of enzymes, it is possible to divide enzymes into those that require free sulfhydryl groups for their action and those that do not. Further information on the function of these groups can be obtained by adding the appropriate substrate before the mercury. If this protects the enzyme, the substrate must be linked in some way with the sulfhydryl group to prevent its combination with the mercury. This occurs in the case of the succinoxidase. If the substrate does not protect, then the sulfhydryl groups are located at other sites in the molecule, and their presence there, like free amino, carboxy, and hydroxyl groups, contributes to the general electronic pattern of the enzyme. This work illustrates how analysis of drug action can increase the understanding of enzyme chemistry. It also offers a logical basis for the development of an antidote to mercury poisoning, which culminated in the discovery of 2,3-dimercaptopropanol * by Peters and his collaborators at Oxford. Apparently the sulfhydryl groups of this compound have a greater affinity for mercury than do the corresponding groups on most enzymes. Mercury therefore combines with it, the enzymes are free to resume their normal function, and if cell dysfunction has not proceeded too far the toxic symptoms disappear.

Conjugation Reactions

It is always somewhat surprising to realize that most drugs are changed when introduced into the body, since many of them bear little chemical resemblance to known metabolites. There are two general mechanisms available to the body to effect this change. First, the so-called detoxication reactions, which have been known for a long time, involving the

---

* This is known as BAL (British anti-lewisite) since it was used in World War I to counteract lewisite, an arsenic-containing poison. It is now used to combat poisoning by mercury and arsenic.

conjugation of the drug with some substance such as glycine, or acetic, glycuronic, or sulfuric acids. The conjugated product may or may not be less toxic than the original drug. These reactions take place mostly in the liver and some of the enzymes which catalyze them are probably there for the express purpose of dealing with potentially toxic compounds taken in with the food or formed by bacterial action in the intestine. The conjugation of phenol with sulfuric acid is an example of this. Any conjugation is a synthesis and, therefore, requires energy. This must be supplied by energy-yielding reactions in the cell, and for this reason the enzymes catalyzing these conjugations have not been clearly defined. In two cases, the conjugation of sulfanilamide with acetic acid and of p-aminobenzoic acid with glycine or acetic acid, it has been possible to purify the drugs partly and to supply the needed energy by the addition of adenosine triphosphate. Table 3, taken from Lipmann,[8]

TABLE 3. CONJUGATION OF 2.5 μM OF VARIOUS
SULFONAMIDES WITH 50 μM ACETATE BY PIGEON
LIVER INCUBATED FOR 20 MINUTES AT 37°

Compound	μM Conjugated
Sulfanilamide	1.09
Sulfathiazole	0.40
Sulfadiazine	0.30
p-Aminobenzoic acid	0.62

shows the relative rates of acetylation of various sulfa drugs and p-aminobenzoic acid. Sulfanilamide reacts most rapidly. It is of interest that substitution on the amide nitrogen reduces the rate of acetylation of the nitrogen on the ring at the other end of the molecule. It is probable that the reactions are catalyzed by the same enzyme which acetylates certain normal metabolites, since pantothenic acid is the coenzyme for both.

When the conjugation of p-aminobenzoic acid with glycine is studied, the conditions are found to be more complex. Cohen and McGilvery[9] showed that the reaction occurs in broken cell suspension of liver, but such preparations must be treated very carefully because freezing or hypo- or hypertonicity causes inactivation. Since these procedures should not affect single enzymes, it is probable that they disrupt an aggregate of enzymes, the orderly arrangement of which is necessary for the synthesis of this peptide bond. It is still problematic whether the reaction can be considered a model for the synthesis of proteins. Studies with labelled amino acids have shown that they can rapidly be incorporated into proteins by liver-cell suspensions under conditions comparable to, but not identical with, those which are optimal for the conjugation of p-aminobenzoic acid.

Hydrolytic and Oxidative Reactions

The second mechanism for dealing with drugs is the actual alteration of the molecule by oxidation, reduction, or hydrolysis. Sometimes such changes are followed by conjugation, but more frequently the product is excreted as such. Four hydrolytic and one oxidative reaction may be considered here to illustrate some of the difficulties of interpretation. One of the older synthetic analgesic drugs is acetanilide. The question in the case of this compound, as well as many others, is whether the active agent is acetanilide itself or some metabolite of it. The formula of acetanilide suggests that hydrolysis into acetic acid and aniline could easily occur, and the aniline formed could then be oxidized. In fact, examination of the urine of individuals who had taken acetanilide showed the presence of some *p*-aminophenol. This compound was tested and shown to have analgesic action. It also produces methemoglobinemia when injected, and methemoglobin is often found in the blood after the administration of acetanilide. The story therefore seemed fairly clear cut. Acetanilide was first hydrolyzed, the aniline was then oxidized to *p*-aminophenol which was probably the active agent. Michel *et al.*[10] were able to show that the livers and kidneys of various animals hydrolyze acetanilide *in vitro*. The enzyme responsible for this is an acylase which hydrolyzes the acetylated amino acids probably normally formed in the body. This identification was based on the fact that the ability of a tissue to hydrolyze acetanilide parallels its ability to hydrolyze acetylated amino acids and also on the identity of the pH optima for the two reactions. When aniline is incubated with tissues, *p*-aminophenol is formed, but this oxidation is catalyzed by a thermostable component, possibly iron or copper salts. The *in vitro* work apparently corroborated the *in vivo* findings. Recent studies by Greenberg and Lester,[11] and Brodie and Axelrod[12] have shown, however, that little free *p*-aminophenol is excreted in the urine after acetanilide is given. Instead, acetanilide is oxidized without previous hydrolysis and the acetyl *p*-aminophenol is then, in part, conjugated with glucuronic acid. It seems, therefore, that hydrolysis of acetanilide which, from the *in vitro* work appeared to be the first step in the main pathway of the metabolism of this compound, is actually but a minor side reaction. This illustrates one of the difficulties mentioned above, namely, that the existence of appropriate enzymes does not necessarily mean their importance for the metabolism of the drug in the body as a whole.

Another difficulty is illustrated by atropine and related compounds. Early workers had shown that rabbits, when given the drug (which is an ester), excreted the hydrolyzed products, tropine and tropic acid. Bernheim and Bernheim[13] obtained various preparations of guinea-pig liver,

rabbit liver, and serum, which hydrolyzed atropine and homatropine but which were almost inactive for scopolamine. The enzyme was closely associated with one which rapidly hydrolyzed simple esters of mandelic acid, the acid in homatropine; but many preparations which hydrolyzed the former had no effect on homatropine or atropine. It has been known for some time that the hydrolyzed compounds are pharmacologically inactive, and the possibility immediately arose that the very great differences exhibited in the tolerance of various species of animals to atropine might be correlated with the presence of more or less enzyme in tissues, which would inactivate the drug. This is, however, not the case. The rat and mouse, for instance, have a higher tolerance than the guinea pig or rabbit, and yet their tissues do not contain the enzyme which hydrolyzes atropine. The presence of the enzyme in the guinea pig and rabbit seems, therefore, to be merely fortuitous. The situation is further complicated by the fact that Glick and Glaubach [14] were able to show, by distribution studies, that separate enzymes exist for atropine, homatropine, cocaine, tropacocaine and scopolamine. It is very probable that the drugs are not the normal substrates for the esterases. But then what is? And why the great specificity? These important questions may be answered as more becomes known about intermediary metabolism. Finally, it may be mentioned here that meperidine, an analgesic ester, is also rapidly hydrolyzed by an esterase quite different from the ones described above. Thus, the *in vitro* studies of the fate of this class of drug have revealed the existence of a number of hitherto unknown highly specific esterases, but they have not thrown much light on the problem they were originally meant to solve, namely, the mechanism of tolerance.

The hydrolysis of hydantoin by a specific enzyme was discovered while looking for a mechanism to account for the metabolism of certain disubstituted hydantoins, such as diphenylhydantoin (dilantin), which are used in epilepsy. When dilantin is administered, some of it is excreted as the corresponding hydantoic acid. In other words, the hydantoin ring is split at what might be termed the peptide linkage. A number of tissue preparations were made to determine whether they contained an enzyme which would catalyze this hydrolysis. No evidence of such was found, and even simpler disubstituted hydantoins, such as the dimethyl compound, remained unchanged. Eventually, as a last resort, the unsubstituted hydantoin was tested and found to be rapidly hydrolyzed to hydantoic acid.[15] The hydantoin enzyme is found in the livers and, in smaller concentrations, in the kidneys of all omnivorous animals, but it is not present in any tissue of the strict herbivores, such as the guinea pig and rabbit. It can be purified and concentrated, and the resulting preparation is apparently specific for hydantoin. Diphenyl hydantoin is not attacked, and the mechanism of its hydrolysis remains obscure. The enzyme in the

body may possibly be able to hydrolyze it slowly, or there may be another enzyme, less stable, the activity of which is lost when the cells are broken and which therefore cannot readily be demonstrated. But what is the hydantoin enzyme doing in liver? The fact that it is only present in meat-eating animals suggests that hydantoin may be a normal constituent of certain animal tissues. The older biochemists thought that hydantoins might be formed from amino acids in intermediary metabolism because they are so easily made in the test tube. But this idea has been abandoned for some time, and hydantoins are not supposed to occur naturally. Lazareff [16] has, however, recently produced some evidence that hydantoin can be formed from glycine and ammonia when these substances are incubated with liver slices. If this is so, the enzyme may have some purpose, and hydantoin formation may be one of the pathways of glycine metabolism. Substituted hydantoins inhibit the enzyme, and so, curiously, do compounds containing the pentamethylene ring, such as metrazol. At first glance this looks like an interesting pharmacological correlation. Metrazol produces attacks which resemble epilepsy, and hydantoins suppress such attacks. The enzyme is not present in the brain, but inasmuch as metrazol and hydantoin compete for the surface of one enzyme, it is not inconceivable that they could also compete for a locus in the brain, with a similar chemical configuration without enzymatic activity. The trouble with this pleasant analogy is that metrazol produces electroencephalographic changes which resemble petit-mal attacks and the hydantoins have no effect on these but suppress only grand-mal seizures. Consequently the two drugs must be acting on different loci in the brain.

A hydantoinase is also found in jack-bean meal, a preparation which contains urease and allantoinase as well. It can be extracted from the meal by slightly alkaline solutions, and allantoinase accompanies it. The two enzymes can be distinguished by the fact that dimethyl hydantoin inhibits the hydrolysis of hydantoin in concentrations which have no effect on the allantoinase system. In every way, the specificity of the jack-bean meal hydantoinase is exactly the same as that of the liver enzyme; it is also inhibited by metrazol. Jack-bean meal hydantoinase has a somewhat more alkaline pH optimum than liver enzyme. Its energy of activation is about twice as high, and it differs from the liver enzyme in its sensitivity to such drugs as cyanide and arsenic. The functions of the enzymes in jack-bean meal are unknown.

No compound looks less like a normal metabolic intermediary than diisopropyl fluorophosphate. Yet, when it is administered to animals it is partly destroyed, and when it is incubated with tissues, especially liver, it is rapidly hydrolyzed into diisopropylphosphate and fluoride ions. Mazur [17] was able to purify the enzyme partially, and it apparently is distinct from any enzyme previously described. Moreover, the enzyme is

found in all the organs studied, including muscle, brain, heart, and intestine. One would therefore assume that it has an important normal function, but what that is cannot even be guessed. Table 4, taken from Mazur, shows the specificity of the enzyme in plasma. The smaller the organic substitution on the phosphate, the more rapid the hydrolysis. Thus, dimethyl fluorophosphate is hydrolyzed about seven times as rapidly as the diisopropyl compound. Curiously, however, the ethyl methyl compound is attacked much more slowly than the diethyl compound. The liver enzyme behaves differently. It hydrolyzes diethyl more rapidly than dimethyl fluorophosphate. It is obvious, therefore, that the enzymes in various tissues differ somewhat in their properties, but this observation is still unexplained.

TABLE 4. HYDROLYSIS OF VARIOUS FLUOROPHOSPHATES IN A CONCENTRATION OF 0.35 μM BY RABBIT PLASMA AND LIVER AT 38° pH 7.4

(Figures are cmm CO_2 Displaced by the Hydrofluoric Acid Formed)

Compound	Plasma	Liver
Dimethyl fluorophosphate	114	73
Diethyl fluorophosphate	77	97
Diisopropyl fluorophosphate	15	22
Ethylmethyl fluorophosphate	14	51

Finally, one oxidation reaction may be described. It has been known for some time that quinine is changed in the body and that it disappears when incubated with various tissue suspensions. Further investigations showed that rabbit tissues contain an enzyme capable of oxidizing quinine to quinine carbostyril. The enzyme is also present in the tissues of the rat, mouse, cat, and guinea pig, but it cannot be demonstrated in other animals. Knox[18] has purified it and finds that its activity parallels that of the aldehyde oxidase, with which it is closely associated, but probably not identical, because the adehyde oxidase is present in the livers of animals which do not oxidize quinine. Like the aldehyde oxidase, the quinine oxidase has flavin adenine dinucleotide as its prosthetic group, and the two enzymes can act as a unit, for when an aldehyde and the carbostyril are added together the former is oxidized and the latter reduced. Besides aldehydes and quinine, other cinchona alkaloids and N-methyl nicotinamide are oxidized by the purified preparation. In fact, cinchonidine is oxidized about seven times as rapidly, and N-methyl nicotinamide about twice as rapidly, as quinine. In mammalian liver preparations, quinidine, the optical isomer of quinine, is oxidized much more slowly than quinine, but Kelsey et al.[19] have shown that the reverse

is true in birds. Here again is an example of an enzyme with a curious specificity and apparently haphazard distribution, the normal function of which may be the oxidation of N-methyl nicotinamide, but which nevertheless oxidizes cinchonidine more rapidly.

In conclusion, a few general statements can be made. Investigations on the metabolism of drugs have revealed a group of enzymes, hitherto unsuspected, to which no known function can yet be assigned. This is even true for the alcohol oxidase, which has been known for a long time, and which accounts for the oxidation of ingested alcohol, although it plays no known part in normal intermediary metabolism. Its presence in the liver, where most of it is, may simply be a protective device for the oxidation of alcohols produced by bacterial action in the intestine. This may also be the function of other enzymes in the liver, which is the main organ for the so-called detoxication reactions. In the other aspect of the problem, the attempt to explain the pharmacological actions of a drug on the basis of its inhibition of certain enzymes, successful correlations are only obtained if the enzyme in question plays a unique and vital part in the functioning of the cell or organism. There are only two such enzymes known at present, the cytochrome oxidase and the cholinesterase, the inhibition of which by certain drugs can be considered an adequate explanation of the effects they produce.

REFERENCES

1. Mazur, A., and Bodansky, O., *J. Biol. Chem.,* **163,** 261 (1946).
2. Jones, Jr., H. W., Meyer, B. J., and Karel, L., *J. Pharmacol.,* **94,** 215 (1948).
3. Bernheim, F., and Bernheim, M. L. C., *J. Pharmacol.,* **57,** 427 (1936).
4. Eadie, G. S., *J. Biol. Chem.,* **138,** 597 (1941).
5. Gross, E. G., Holland, H., Carter, H. R., and Christensen, E. M., *Anesthesiology,* **9,** 459 (1948).
6. Kuhn, H. H., and Surles, D., *Arch. intern. pharmacodynamie,* **58,** 88 (1938).
7. Barron, E. S. G., and Singer, T. P., *Science,* **97,** 356 (1943).
8. Lipmann, F., *J. Biol. Chem.,* **160,** 173 (1945).
9. Cohen, P. P., and McGilvery, R. W., *J. Biol. Chem.,* **169,** 119 (1947).
10. Michel, H. O., Bernheim, F., and Bernheim, M. L. C., *J. Pharmacol.,* **61,** 321 (1937).
11. Greenberg, L. A., and Lester, D., *J. Pharmacol.,* **90,** 68 (1947).
12. Brodie, B. B., and Axelrod, J., *J. Pharmacol.,* **94,** 29 (1948).
13. Bernheim, F., and Bernheim, M. L. C., *J. Pharmacol.,* **64,** 209 (1938).
14. Glick, D., and Glaubach, S., *J. Gen. Physiol.,* **25,** 197 (1941).
15. Bernheim, F., and Bernheim, M. L. C., *J. Biol. Chem.,* **163,** 683 (1946).
16. Lazareff, A., *Doklady Akad. Nauk. S.S.S.R.,* **61,** 677 (1940).
17. Mazur, A., *J. Biol. Chem.,* **164,** 271 (1946).
18. Knox, W. E., *J. Biol. Chem.,* **163,** 699 (1946).
19. Kelsey, F. E., Oldham, F. K., and Geiling, E. M. K., *J. Pharmacol.,* **85,** 70 (1945).

GENES AND BIOCHEMICAL REACTIONS

David M. Bonner

Osborn Botanical Laboratory, Yale University, New Haven, Conn.

The Gene and the Problem of Gene Action

By definition, the smallest unit of inheritance is the gene. Specific genes are associated with specific hereditary traits. The ingenious work of many investigators during the past three decades, using organisms such as the fruit fly (*Drosophila*) and the corn plant, has provided a wealth of information concerning the properties of these units. Two characteristic properties of genes are self-duplicability and mutability. A gene is able to cause the production of a unit exactly like itself, as indeed it does at each cell division. While genes are on the whole remarkably stable, they can undergo change, or mutation, and the altered genes are in turn duplicated. Indeed, it is this property of mutability which has enabled the characterization of genes, since a gene can be recognized only after a change has occurred with a concomitant alteration of function. Genes are contained in chromosomes, in linear order, the chromosomes being contained within the cell nucleus. Thus the chromosomes, and in turn the genes, are separated from the cytoplasmic cellular constituents by the nuclear membrane. Cells may be diploid, as in the fruit fly, and contain duplicate chromosomes and consequently duplicate genes, or they may be haploid, as in many fungi, and contain but a single set of chromosomes. A great many genes are now known in a variety of organisms, and indeed the relative positions (loci) and in some instances the absolute positions of the genes on the chromosomes have been determined.

Since formulation of the gene theory geneticists have been faced with the general problem of the mechanism of genetic control. Genes have long been known to control morphological characters, but how this control is exercised still remains a major problem in genetics. Within the past decade very convincing evidence has been found for the idea that genes determine what biochemical reactions occur within a cell. It is in this field that geneticists and biochemists have collaborated in an en-

deavor to penetrate more deeply into the relationship of gene to bio-
chemistry. The term biochemical genetics has been applied to this aspect
of genetics.

During the past twenty years a considerable body of evidence has been
accumulated suggesting that genes exert their morphological control
through control of cellular biochemical reactions. Biochemical and ge-
netic investigations of eye color inheritance in the fruit fly and of pig-
ment inheritance in plants (see review by Beadle [3]) suggest the view that
specific genes determine the occurrence of specific biochemical reactions.
Indeed, some of the earliest noted evidence of this relationship is to be
found in inborn errors of metabolism in man, as described by Garrod.[18]
A study of the genealogy of families afflicted with such diseases as alcapto-
nuria and phenyl oligophrenia clearly demonstrates that loss of a specific
gene gives rise to loss of a specific biochemical function. Because of the
many obvious difficulties in the use of insects, higher plants, or mammals
for a critical investigation of the relationship of genes to biochemical
reactions, attention has turned to microorganisms.

Microorganisms such as bacteria, yeasts and fungi provide excellent
material for this type of investigation, since their life cycles are short
and many can be grown on a chemically-defined medium. The organism
which was first used for this type of work and has been most extensively
studied is the fungus, *Neurospora crassa*. The work of Dodge [16] and of
Lindegren [30] had established the usefulness of this organism as a genetic
tool, and Beadle and Tatum [4] therefore selected it for a critical study of
the genetic control of biochemical reactions.

Neurospora as a Tool in Biochemical Genetics

The majority of the work dealing with biochemical genetics of *Neuro-
spora* has been carried out on *Neurospora crassa*, a heterothallic asco-
mycete. The life cycle of this organism is shown in Figure 1. The or-
ganism grows ordinarily as branched hyphae. These vegetative hyphae
give rise to orange-colored, multinucleate spores known as conidia, which
when transferred to fresh medium give rise again to the original vege-
tative strain. These vegetative strains are haploid. The diploid stage of
the organism is transient and is brought about by fusion of a conidium
of one mating type with a protoperithecium of opposite mating type.
The two mating types appear identical and can be distinguished only by
the fact that when the two strains are grown together fruiting bodies are
formed. The fusion nucleus, resulting from the fusion of the two nuclei
of opposite mating type, undergoes two meiotic divisions and one mitotic
division, thereby producing eight haploid ascospores (Figure 1). These
spores, after activation, give rise again to haploid vegetative cultures.
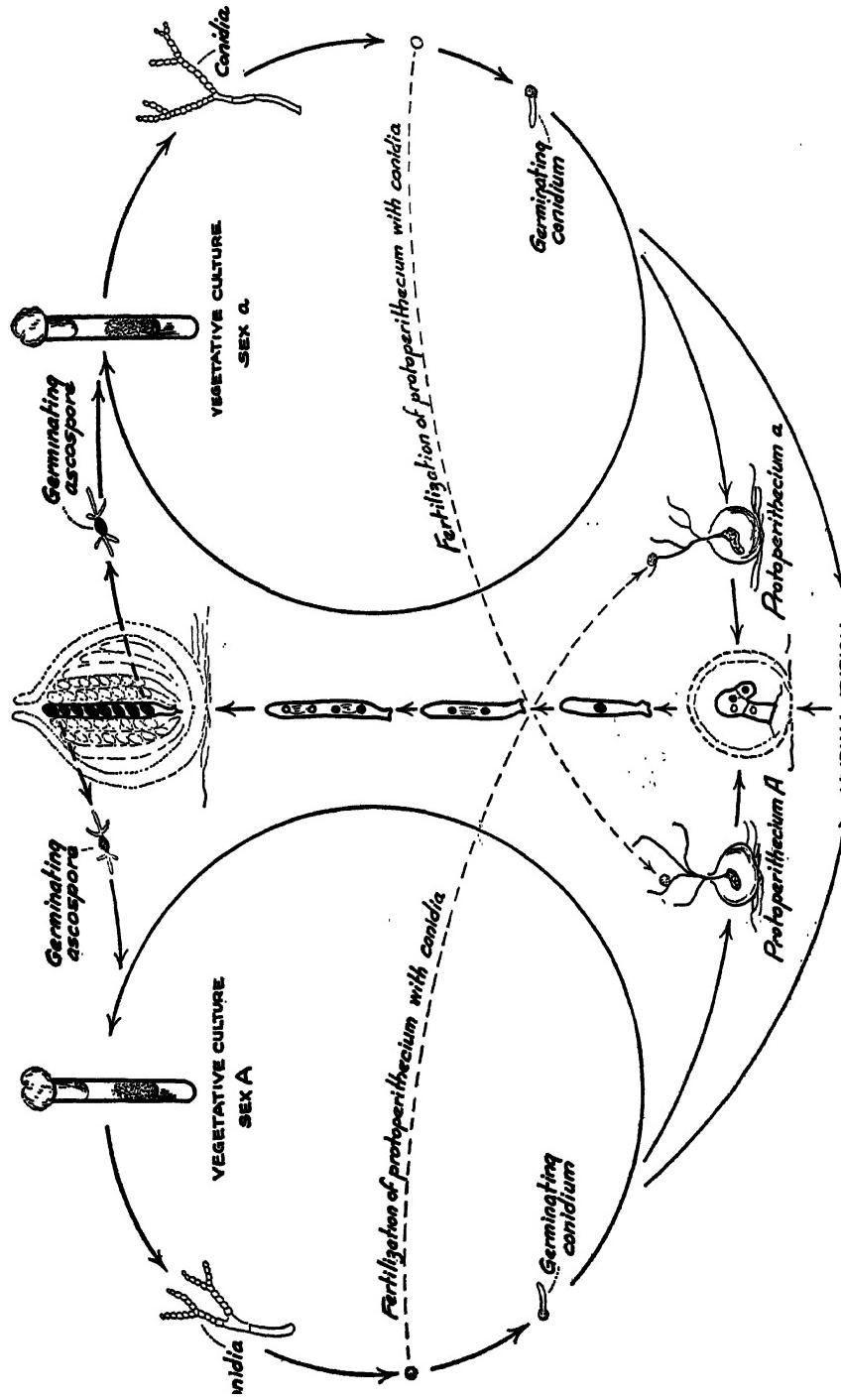
Courtesy "Science in Progress," (Yale University Press, New Haven, Conn.)

Figure 1. Life cycle of Neurospora.

Thus, from a genetic standpoint this organism is exceptionally easy to work with, since all the products of meiosis are preserved. The fact that the vegetative strains are haploid is of great aid since problems of dominance are of no concern.

Neurospora will grow on a chemically-defined medium. It requires for growth a carbon source, such as sucrose, a nitrogen source, such as ammonium nitrate, inorganic salts, and the one vitamin, biotin. If genes do indeed control biochemical reactions, it was argued that it should be possible to induce gene mutations affecting the synthesis of such vital cellular components as vitamins and amino acids. Such mutations could be detected by an increased complexity of growth factor requirements,
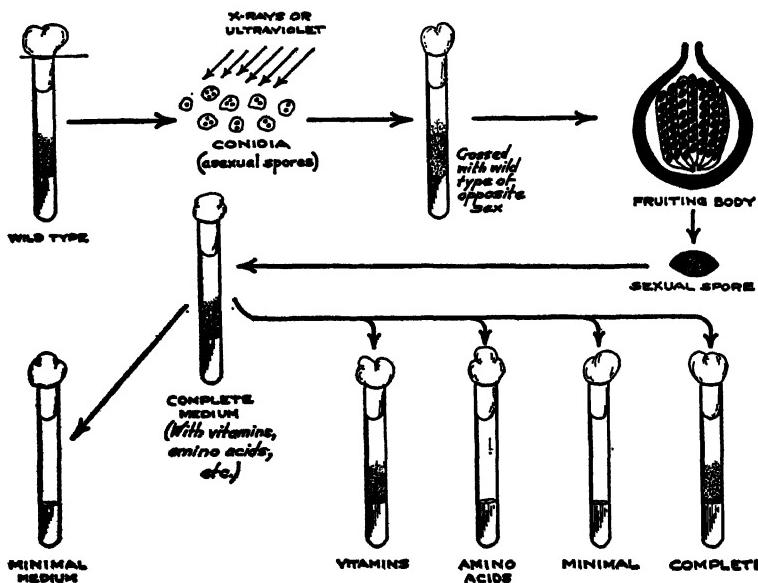
Courtesy "Science in Progress," (Yale University Press, New Haven, Conn.)

Figure 2. Experimental procedure by which biochemical mutants are produced and detected in *Neurospora.*

since mutant strains would be dependent on an exogenous supply of these compounds. Beadle and Tatum[4] accordingly carried out the experiment shown in Figure 2. Conidia were treated with ultraviolet light or x-rays, agents known to increase mutation rate in other organisms. Since conidia are multinucleate, these treated spores were crossed with the opposite mating type of the parental strain to permit the segregation of genetic characters. Single ascospores were selected from this cross. To insure growth of spores having increased nutritional requirements, the spores were established on a complex medium containing yeast extract and hydrolyzed casein as sources of additional growth factors. These

isolated strains were next tested for their ability to grow on a simple medium capable of supporting the growth of the parental strain, i.e., a medium containing only sucrose, nitrate, inorganic salts and biotin. Most of the isolated strains grew equally well on both media, but an occasional strain appeared which grew well on the complex medium and
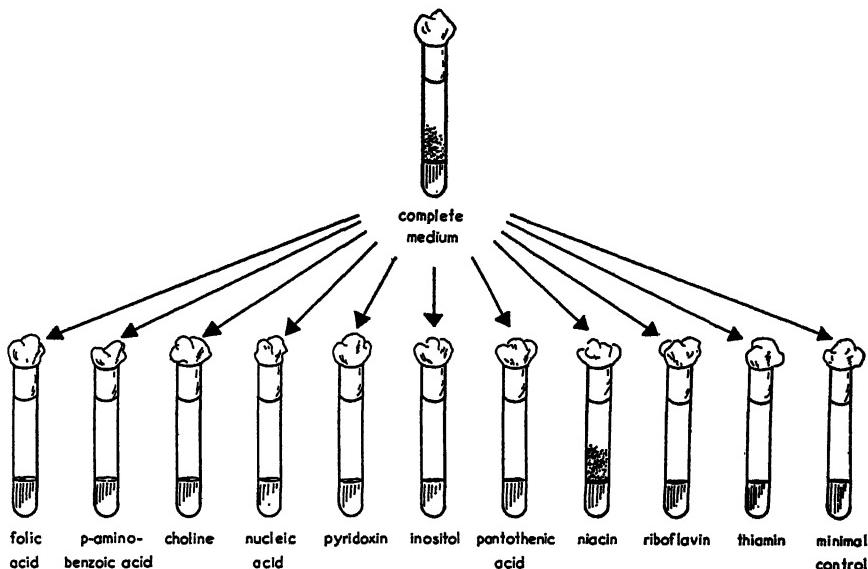
Figure 3. Tests of mutant strain on individual vitamins. Since purines and pyramidines are often found to be required by mutant strains, hydrolyzed nucleic acid is included.

not at all on the minimal synthetic medium. The series of tests shown in Figures 2 and 3 was then used to determine the factors necessary for the growth of such strains. In the illustration, the vitamin, niacin, is the additional factor required. Thus, following treatment with a mutagenic agent it is possible to isolate strains which differ from the parental strain by an added growth-factor requirement.

Gene Control of Biochemical Reactions

Genetic Studies. The fact that strains with increased growth-factor requirements result from treatment with mutagenic agents does not in itself prove that they result from gene mutations. It is possible, however, to test for the genetic basis of these requirements. If one crosses the mutant strain requiring niacin with the opposite mating type of the parental strain, Mendelian segregation of the mutant genes should occur. If the eight ascospores contained in a single ascus of the mature fruiting body resulting from such a cross are dissected out in order and ger-

minated (Figure 4), the resulting eight cultures will all grow on a medium containing niacin, while only four will grow on a medium lacking niacin. Thus four of the cultures require niacin for growth and four are niacin-independent. The requirement for niacin segregates as a one-one ratio as would be expected in the segregation of a single gene difference. The niacin requirement, therefore, results from a single gene difference between parental type and mutant strain.

During the past eight years mutations involving nearly all of the known vitamins and amino acids have been found. Also, mutations affecting pyrimidine and purine synthesis, carbohydrate utilization, and nitrate
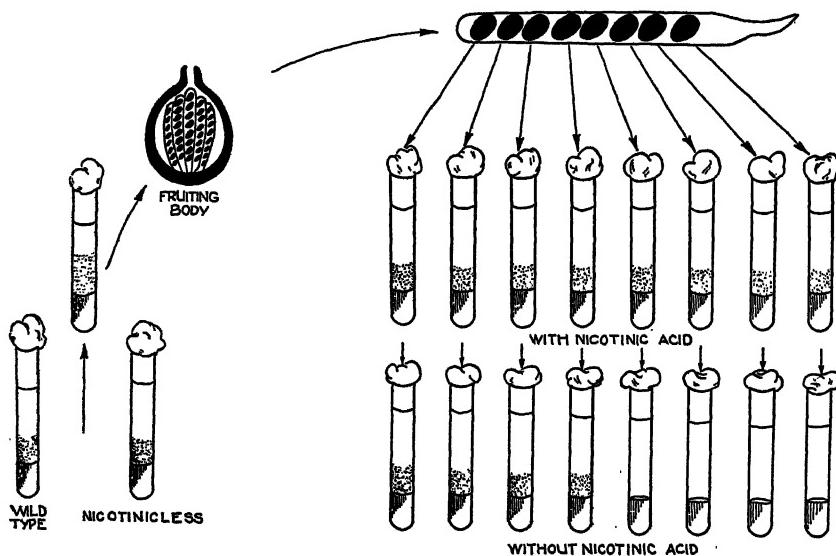
Figure 4. Scheme by which the inheritance of a mutant type is determined.

reduction have been found. Gene control is therefore not limited to niacin synthesis, but is known for a great many vital cellular constituents. The fact that mutations affecting the synthesis of vitamins and amino acids have been found most frequently is a reflection of existing detection methods. Probably the syntheses of all cellular compounds, including large molecules, are gene-controlled, but we are still limited to the detection of gene alterations affecting the synthesis of compounds that can permeate the plant cell wall.

Biochemical Studies (*Tryptophan-Niacin Synthesis*). Among the existing mutant strains of Neurospora, several have been found to require the vitamin, niacin (3-carboxy pyridine).[11] All of these strains, when crossed with the parental strain, are found to differ from it by alteration of a single gene. It is of interest, however, to determine whether the same gene

is affected in all of them or whether alteration of any one of several genes may lead to niacin deficiency. Genetic differences may be determined by crossing the niacin-requiring strains to each other. If the same gene is affected in each mutant, all of the spores resulting from the cross are niacin-dependent. If different genes are affected, some parental-type niacin-independent spores result from recombination. When this experiment is carried out with the niacin-requiring strains, four separate genes are found essential for the synthesis of niacin. Alteration of any one of these four genes gives rise to a niacin-dependent culture.

The material which has just been reviewed shows simply that loss of a specific gene results in loss of a particular synthetic capacity. Let us now examine this genetic loss in its relation to biochemical differences. Genetic analysis of the niacin-requiring strains reveals at least four different genes necessary for niacin synthesis. Biochemcial analysis of these strains likewise reveals four distinct biochemical classes. If one knows many of the biological precursors of a given vitamin or amino acid, the biochemical differentiation of various mutant strains is greatly simplified. Such simplification has been possible in a few cases, as, for instance, in the differentiation of mutant strains requiring arginine [43] or those requiring methionine.[22] As a rule, however, our inadequate knowledge of the biosynthesis of vitamins and amino acids has forced a different approach to biochemical differentiation. Since each of these four niacin-requiring strains is blocked in some step essential for niacin synthesis, accumulation of the normal intermediate may occur and can be revealed by cross-testing culture filtrates (culture medium after growth of the strain) of these four strains to determine whether these culture filtrates have niacin activity for one or more of the niacin-requiring strains. The result of such an experiment is shown in Table 1.

TABLE 1. CULTURE FILTRATES FROM NIACIN-REQUIRING STRAINS OF NEUROSPORA, TESTED FOR THE ACCUMULATION OF NIACIN INTERMEDIATES

Tested with Strain Type	Culture Filtrate from Strain Type			
	1	2	3	4
1	—	—	—	—
2	—	—	—	—
3	—	+	—	—
4	—	+	+	—

— indicates no growth of test strain
+ indicates growth of test strain

From Table 1 we see that culture filtrates of type one have no niacin activity for any of the strains. Culture filtrates of type two, however, will support the growth of types three and four, and those of type three will support growth of type four. From these data various conclusions can be drawn. Since these culture filtrates will not support growth of the strain

from which they were obtained, and since the niacin requirement for all of these strains is nearly identical, the activity of these culture filtrates must represent compounds which can serve as niacin precursors. One can also conclude that these four genetic types represent four distinct biochemical classes. Types one and two grow on no culture filtrates and hence must differ from strains three and four. Type three differs from type four since type three accumulates a substance having niacin activity for type four. Types one and two must differ biochemically from each other since type two accumulates a substance with niacin activity for strains three and four and type one does not.

Assuming a sequential series of reactions to be characteristic of niacin synthesis, the order of genetic blocks must be as follows:

$$W \xrightarrow{\text{type 4}} X \xrightarrow{\text{type 3}} Y \xrightarrow{\text{type 2}} Z \xrightarrow{\text{type 1}} Niacin$$
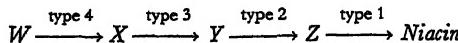
Most of the various compounds in the culture filtrates showing niacin activity have now been identified, and it is possible to give the scheme in chemical terms. Krehl et al.[26] first noted that the amino acid tryptophan can replace niacin in niacin-deficient rats. Strains of type four were likewise found to be capable of growing on either niacin or tryptophan.[5] Since the other niacin-dependent genetic types cannot utilize tryptophan, this one genetic type is clearly differentiated biochemically. The compound accumulated by type two strains was found to be 3-hydroxyanthranilic acid,[9, 35] while the niacin activity of culture filtrates obtained from strains of type three is found to be α-acetylkynurenine.[51] Niacin synthesis appears, therefore, to proceed from tryptophan with the formation of 3-hydroxyanthranilic acid. The mechanism of conversion of 3-hydroxyanthranilic acid to niacin has only recently been clarified. The fact that rats given large quantities of tryptophan excrete quinolinic acid (2,3-dicarboxypyridine) in their urine [41, 19] suggests that this compound is concerned in the conversion of tryptophan to niacin. Quinolinic acid has slight activity for Neurospora,[9, 35] yet mutant strains of type one were found to accumulate this compound in the culture medium.[12] This accumulation suggests that the conversion of 3-hydroxyanthranilic acid to niacin occurs by cleavage of 3-hydroxyanthranilic acid in the 3,4-position to form an aliphatic compound. This substance, after decarboxylation, undergoes ring closure. If decarboxylation cannot occur, ring closure gives quinolinic acid. One can, therefore, formulate the following scheme of niacin synthesis in Neurospora, as shown in Figure 5.

This is a general scheme, and no attempt has been made to include the detailed process of the conversion of tryptophan to 3-hydroxyanthranilic acid. Kynurenine and 3-hydroxykynurenine are thought to serve as inter-

mediates in this conversion,[35] and indeed some evidence is at hand suggesting that the role of tryptophan in niacin synthesis may be indirect.[9] However, the scheme presented in Figure 5 points up clearly several significant facts relating to biochemical genetics. This scheme shows clearly that each of the four genes known to affect niacin synthesis controls a different step in the synthesis. Each gene may be presumed to affect a single specific step. Loss of gene one, for instance, leaves the organism unable to decarboxylate the open-chain intermediate and leads to production of quinolinic acid. Indeed, in many other cases we know the specific reaction which is lost as the result of a specific gene mutation.
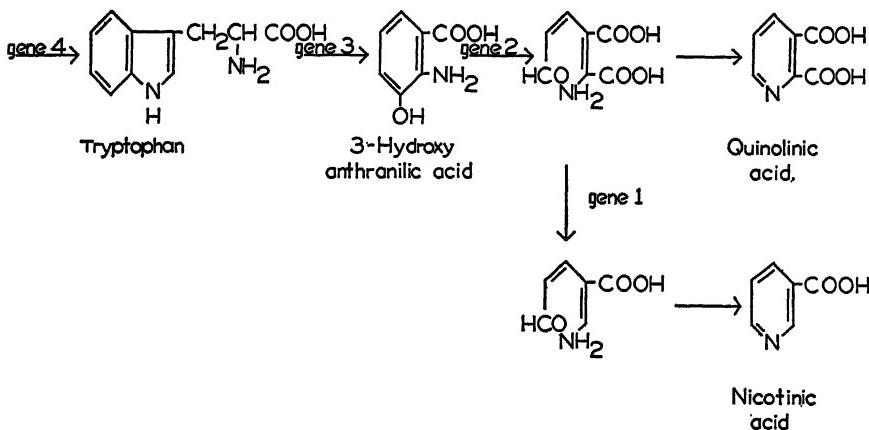
Figure 5. Scheme of niacin synthesis. Genetic blocks are indicated by gene numbers.

Our inability to define precisely the biochemical step blocked, as we cannot define that blocked by gene four, results simply from our inadequate knowledge of cellular biochemistry. As our knowledge in this field grows, such specific defined reactions will also doubtless be found to be affected by specific mutations.

Biochemical Studies (*Isoleucine-Valine Synthesis*). It should be pointed out that gene control of biochemical reactions is not limited to niacin synthesis. Genetic analysis has not been carried out for all of the known nutritional mutants, though it has been for a great many. Each nutritional deficiency which has been examined genetically has been found to result from a single gene alteration. It has also been observed that, in general, several genes are concerned with the synthesis of each vitamin or amino acid. Single gene mutations involved in the synthesis of such other compounds as choline,[21, 23] arginine,[43] and methionine[22] are also correlated with the loss of single known reactions. Biochemical investigations, in short, show that a gene mutation which gives rise to a growth-factor requirement represents loss of ability to carry out a specific biochemical reaction.

The series of experiments which were outlined above points up still other problems. A mutant strain requiring an added growth factor is not biochemically identical with the parental strain, even though its growth rate, when it is grown on exogenous supplies of the factor in question, is normal. One must reckon with the disturbed biochemical balance occasioned by the loss of ability to perform a particular reaction and the resulting accumulation of a precursor prior to the point of block. This accumulation may have no apparent ill effect, as in the case of 3-hydroxy-anthranilic acid accumulation. The precursor may be metabolized in an abnormal manner, as in the accumulation of quinolinic acid. It may also compete with other biochemical reactions and lead to an apparent double requirement. This situation is most strikingly shown by mutant strains requiring isoleucine and valine. A mutant strain is known which differs from the parental strain in a single gene and requires the two amino acids, isoleucine and valine.[18] Detailed investigation has shown that this double requirement is not the result of a common biosynthetic pathway. Rather it appears that genetic loss of isoleucine synthesis leads indirectly to a block in valine synthesis. Experiments using the synthetic keto acid analogs of the two amino acids show that ketoisoleucine strongly inhibits the amination of ketovaline.[8] Since the mutant is unable to utilize ketoisoleucine in the presence of valine, but can utilize keto-valine in the presence of isoleucine, the obvious interpretation is that because of this genetic block the keto acid analog of isoleucine accumulates and competitively inhibits the amination of the keto acid analog of valine. Thus the genetic block in one biosynthetic step in this instance results in an apparent double requirement. This mutant has recently been shown to accumulate the dihydroxy analog of isoleucine.[1] The relationship of this compound to ketoisoleucine has not been established. Nevertheless, the experiments using synthetic keto acid analogs emphasize the importance of metabolic antagonisms occurring as the result of accumulated intermediates. In a critical biochemical evaluation of such mutant strains, the metabolic disturbances that may be occasioned by loss of biochemical function must be carefully considered.

Enzyme Studies. One may now say with a certain degree of assurance that genes control biochemical reactions. This in itself, however, tells little concerning the actual mechanism involved. It has often been postulated that genic control is exercised through control of enzyme specificity. Since nearly all biochemical reactions require enzymatic catalysis, the assumption that genes control biochemical reactions through control of enzyme specificity is attractive and plausible. Direct evidence in support of this view, however, remains scant. If genic alteration gives rise to enzyme alteration, one might expect substrate accumulation. Indeed, this has been noted in several instances. Accumulation of 3-hydroxyanthra-

nilic acid is an example. This evidence is indirect and the isolation of the altered enzyme would be more convincing. One instance studied in detail deals with the conversion of indole to tryptophan. Indole has been shown to be converted to tryptophan in *Neurospora* in a single step by condensation with the amino acid serine.[45] Umbreit *et al.*[47] studied the enzyme involved in this condensation and found that it required the presence of pyridoxal phosphate. Recently Mitchell and Lein [34] obtained a mutant blocked in the conversion of indole to tryptophan and reported that the enzyme described by Umbreit *et al.* was lacking. This instance shows that alteration of a specific gene causes loss of ability to carry out a specific biochemical reaction, which in turn is associated with loss of activity of a specific enzyme.

. Similar observations have been made regarding the enzyme causing the hydrolytic splitting of lactose.[10] Wagner,[48] however, has found that a genetic block of a specific reaction is not always associated with loss of a specific enzyme. In investigations of certain pantothenic-less strains, he finds that a cell-free extract of the parental strain can couple β-alanine and pantoyl lactone with the formation of pantothenic acid.[40] Certain mutant strains are known to require pantothenic acid, and of these, one has been shown to be unable to grow if furnished β-alanine and pantoyl lactone; yet it does synthesize these two compounds.[46] Wagner has recently shown that extracts of this mutant strain also can couple the component parts of pantothenic acid.[48] Thus we have, in this instance, alteration of a specific gene causing loss of ability to carry out a specific reaction without loss of the specific enzyme. Loss of function in this instance arises presumably from inhibition of the enzyme system, the inhibition in turn resulting from gene alteration. This case may well represent a situation analogous to that described for strains requiring isoleucine and valine.

Is *Neurospora* Unique? Before entering into a discussion of possible interpretations of the material just reviewed let us examine briefly the possibility that the phenomena encountered in *Neurospora* are peculiar to this organism. Biochemical mutations have now been induced in a variety of microorganisms including several fungi, bacteria and yeast. Many of these organisms have no known perfect stage and hence are not adapted to genetic analysis. In these forms the mutant strains closely resemble the *Neurospora* strains.[7, 14] Genetic studies of biochemical mutants of *Saccharomyces cerevisiae*,[39] *Ophiostoma multiannulatum*,[17] and *Ustilago maydis*,[38] point clearly to the genetic control of biochemical reactions and suggest that this control is similar to that in *Neurospora*. Strain K-12 of *Escherichia coli* shows a similar relation of gene and biochemical reaction. Lederberg and Tatum [29] have shown that genetic recombination of biochemical characters occurs in this organism and have devised techniques enabling them to work with this strain genet-

ically.[27] *E. coli* genetic control of biochemical reactions is known to occur, and again the relationship of gene to biochemical reaction appears essentially similar to that found in *Neurospora*. As mentioned in the initial part of this paper, the study of inborn errors in man leads to the conclusion that the same relationship obtains in man. Indeed, we have no reason to assume that the mechanism of the gene control of biochemical reactions is substantially different throughout the living world.

Gene-Enzyme Relationships

The observations which have been reviewed lead to the conclusion that genes control biochemical reactions, and that this control is probably exercised through control of enzyme production. The fact that single genes control single biochemical reactions has led to the hypothesis that a one-one relationship exists between gene and biochemical reaction.[24] Whether this one-one relationship of gene and biochemical reaction is a reflection of a more fundamental relationship existing between gene and enzyme, i.e., whether single genes determine the specificity of single enzymes, will be discussed later. The problem of the genetic control of biochemical reactions, and perhaps of genetic control in general, becomes then primarily the problem of the gene control of specific enzyme formation.

Detailed information on the chemical composition of the gene is not available. It is known, however, that chromosomes are composed of nucleoprotein.[33] Since chromosomes contain nongenic material, however, this in itself does not constitute proof of the nucleoprotein nature of the gene. Other lines of evidence contribute to the supposition that genes are composed of nucleoprotein. Ultraviolet light induces gene mutation. If one compares the efficiency of ultraviolet light on inducing gene mutations at various wavelengths, one finds maximum efficiency in the region of the absorption maximum of nucleic acids, suggesting that nucleic acids are genic constituents. One may also reason from the similar properties of viruses and genes that genes are nucleoproteins, as are certain viruses.[44] The important similarity of genes and viruses for the purpose of this discussion is that they both possess the property of self-duplication. A gene possesses the ability to catalyze the production of a unit exactly like itself and hence may be considered as having enzymatic properties. A gene appears, therefore, to contain nucleoprotein, to have autocatalytic properties, and to determine the enzymatic structure of the cell.

Gene Action

Having reviewed the evidence concerning the genetic control of biochemical reactions, and having considered briefly the properties of the

gene, let us examine the problem of gene action. As mentioned earlier a one-one relation has been observed between gene and biochemical reaction. This relationship suggests that the specificity of each of the cytoplasmic enzymes is directly gene-determined. The relationship between gene and biochemical reaction in such a concept is shown diagrammatically, with G denoting gene, P, enzyme precursor, E, any cytoplasmic enzyme, and $A \rightarrow B$, a cellular biochemical reaction.

$$
\begin{array}{ccc}
 & P & \\
G \longrightarrow & \downarrow & A \\
E \longrightarrow & & \downarrow \\
 & & B
\end{array}
$$
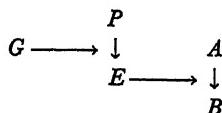
The observation that a single gene change results in loss of ability to carry out a single biochemical reaction is in general agreement with such a theory. The agreement is perhaps even more clearly shown in the observation of Mitchell and Lein [34] that loss of ability to condense indole and serine in the formation of tryptophan is correlated with loss of enzyme activity. These observations, however, cannot be considered as critical proof of such a thesis since they are subject to other interpretations. Wagner's observations,[48, 49] discussed earlier, superficially disagree with this interpretation, though the factor responsible for inactivity of the enzyme system must be worked out before critical evaluation of these results can be made. Irreconcilable with such a thesis, however, are certain observations made on *E. coli*. Lederberg [28] has reported that any one of seven independent genes affects the production of the enzyme lactase in *E. coli*. It is probable that alteration of either of two independent genes in *Neurospora* results in defective lactase production.[10] Thus in these instances more than one gene determines the specificity of a single enzyme.

The observation that loss of any one of several genes leads to loss of lactase synthesis suggests that enzymes, like other cellular constituents are synthesized by a sequential series of characteristic reactions. The specificity of some enzymes appears, therefore, to be determined by more than a single gene, and a concept demanding that the specificity of each cellular enzyme is controlled by a single gene proves unsatisfactory. That single genes may control single biochemical reactions tells little concerning the precise mechanism involved.

Since genes are involved in enzyme synthesis, the mechanism of enzyme formation must be considered. Unfortunately we know relatively little concerning the synthesis of enzymes. As just pointed out, genetic work suggests that enzymes, like other cellular constituents, are synthesized by a sequential series of characteristic biochemical reactions, under genic control. Recent work by Spiegelman [42] and Monod [36] suggests that each

enzyme is not necessarily synthesized *de novo* from constituent parts, but rather may be synthesized at the expense of other enzymes. Spiegelman,[42] in fact, has very reasonably suggested that many enzymes may be made from a common enzyme precursor pool. Monod's observations that certain pairs of enzymes show the phenomenon of diauxie with each other might further suggest that many such enzyme precursor pools are involved in the synthesis of the various cytoplasmic enzymes. Whether the enzyme precursors represent large complex molecules or not cannot, of course, be decided at present. It seems logical, however, to assume that the molecules of the enzyme precursor pools are large, since the specific enzymes themselves are large molecules. If this were the case, one might expect that the number of biochemical steps involved in the synthesis of a specific enzyme from the precursor pool would be few, indeed perhaps a single step in many cases. These enzyme precursors, by definition, do not have the specificity of the final enzyme, and their synthesis might be expected to be genetically similar to the other biochemical reactions previously discussed, that is, the enzymes involved in the synthesis of these enzyme precursors need not be directly gene-determined. It would seem logical to assume, however, that since the gene itself is a specificity factor, the gene is fundamentally involved in the process of conversion of enzyme precursor to specific enzyme.

These considerations lead us finally to the problem of the conversion of enzyme precursor to specific enzyme. The individual steps in the reaction chain from enzyme precursor to specific enzyme may be enzymatically controlled, but the specificity of these enzymes controlling the conversion of precursor to specific enzyme must be gene-determined. In fact, it may be that only those enzymes involved in the conversion of enzyme precursor to specific enzyme are under direct genic control. The specificity of these enzymes in relation to the gene must nevertheless be explained. Thus, fundamentally, the problem of gene action appears directly concerned with the determination of the specificity of certain enzymes. Do the genes themselves catalyze the reactions involved in the formation of specificity or do gene products bring it about? We have virtually no evidence permitting us to favor one point of view over the other. Since genes are nucleoprotein, one might expect that each gene possesses a characteristic spacial configuration. The plausible suggestion has, therefore, been made that genes serve as templates in determining the configuration of enzymes. This interpretation differs from the concept previously discussed in that it makes provision for more than a single specific step in the formation of a specific enzyme. Such a thesis, however, has serious obstacles to overcome. It implies tremendous passage of protein through the nuclear membrane in a regulated fashion. While this passage of protein is not impossible, it is difficult to visualize in our present state of knowledge.

In part to obviate this difficulty, a great deal of attention has recently been given to gene-initiated plasmagenes.[42] Such particles were considered as direct gene products which moved to the cytoplasm, retaining the self-duplicating property of the gene. These plasmagenes, in turn, were thought to mediate the conversion of enzyme precursor to specific enzyme. No data are available in support of the above concept. It is equally logical to suppose that some other particle, one lacking self-duplicating properties, represents the initial gene product, which in turn might mediate specific enzyme formation. Nucleic acid might be such a molecule. Genes are known to contain nucleic acid, and bacterial transformations, which perhaps should be considered as directed mutations, are known to involve nucleic acid as the specific transforming principle.[2, 6, 32] Nucleic acid is thought to be involved in protein synthesis also.[15] It is equally reasonable, however, to assume that the initial gene product is a small molecule.[50] This small molecule might unite with an enzyme precursor in the cytoplasm and form an enzyme, which in turn could catalyze the formation of specific enzyme. Diagrammatically such a relationship might be shown as follows, with G denoting gene, GP, gene product, EP, enzyme precursor, E, specific enzyme, and $A \rightarrow B$ a specific cellular biochemical reaction.

$$G \longrightarrow GP \longrightarrow EP$$
$$\downarrow$$
$$E$$
$$\downarrow$$
$$A \longrightarrow B$$
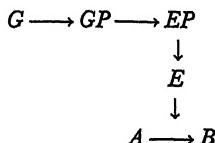
This interpretation has several merits. It predicts that while the correlation of gene to cellular enzyme may be one it is not invariably one, since several steps may be involved in enzyme formation. Indeed several steps have been experimentally observed. It also obviates the difficulty encountered in the passage of generous amounts of large molecules through the nuclear membrane. However, this interpretation, like the others, cannot be critically tested without much more detailed knowledge concerning the synthesis of enzymes.

The latter remarks have been largely speculative, since our knowledge of the structure of proteins and the synthesis of enzymes is too inadequate to permit more than speculation. They have been an attempt to explain logically the meager facts at our disposal. How might one then summarize current thought regarding the mechanism of gene action? The material reviewed demonstrates clearly the genetic control of biochemical reactions. Indeed, there is a striking correlation between alteration of a specific gene and loss of a specific biochemical reaction. There is also a correlation between enzyme production and gene. That single genes

determine directly the specificity of all single enzymes seems unlikely. What the precise relation is, however, between gene and enzyme remains a major biological enigma.

Evolutionary Considerations

In addition to the problem of gene action, investigations in the field of biochemical genetics have called attention to certain evolutionary problems. The investigations of Lwoff,[31] Knight,[25] Schopfer,[40] and others, on the growth-factor requirements of many different microorganisms, suggested several years ago the general principle that evolutionary specialization with regard to nutritional requirements was accompanied by loss of synthetic ability. The biochemical and genetic work on *Neurospora,* bacteria, and yeast furnishes a genetic basis for such evolutionary specialization. Starting with a completely autotrophic organism, one can readily see how its nutritional requirements would become increasingly complex as gene mutations occurred, giving rise to loss of synthetic ability. Indeed, one might imagine organisms such as viruses and bacteriophage representing the ultimate in this specialization.

The reverse type of evolution, however, is much more difficult to understand; yet from a mechanistic standpoint such evolution must have occurred in past geologic periods. Biochemical investigations of mutant strains of microorganisms suggest that vital cellular components, such as vitamins and amino acids are synthesized by characteristic sequential series of reactions. The reaction chains consist of several steps, and the intermediates involved are apparently of value to the organism only in terms of the specific end product. Thus one is faced, in accounting for the evolutionary development of such reaction chains, with the problem of explaining the acquisition of individual synthetic abilities, which confer no selective advantage on the organism, unless it acquires the entire reaction chain. If the evolution of ability to synthesize an end product such as niacin depends on the simultaneous mutation of all four genes known to control its synthesis, its occurrence seems improbable. The universal occurrence of such compounds as vitamins and amino acids throughout the living world places them among the essential components of living matter. The widespread occurrence of similar reaction chains characteristic of the synthesis of a given vitamin or amino acid further suggests that the biosynthetic mechanisms involved in the synthesis of these compounds must have evolved early in the history of living matter. Oparin [37] has suggested that complex organic matter was synthesized during the cooling of the earth. This organic matter might have accumulated, since at this period there was no microbial life to metabolize it to simpler compounds. If one accepts this thesis, together

with the thesis that the first self-duplicating units of living matter reproduced by taking up complex organic molecules from a medium in which the organic molecules had accumulated, then, as Horowitz [20] has reasonably suggested, these reaction chains might well have evolved backwards from the end product to an early intermediate. Thus a self-duplicating unit, initially dependent on end product for reproduction, would have a strong selective advantage conferred upon it if it acquired the ability to convert an intermediate to the end product. This concept affords a mechanism for the original rapid acquisition of the reaction chains characteristic of the synthesis of vital cellular components. In the world of today this mechanism would supposedly play a minor evolutionary role compared to specialization by loss, since the concept of backward evolution demands the presence of the complex end products and intermediates. One would not expect the accumulation of such compounds at present, since they would be speedily metabolized by microbial activity.

Conclusion

In this discussion of biochemical genetics the relationship of gene to morphology has been omitted. The science of genetics is largely based on the genetic control of morphological characters. Morphological differences probably represent poorly understood differences in cellular biochemistry. The genetic study of such characters has revealed many complicated genetic phenomena. At the present time one cannot account satisfactorily for all known genetic relationships, such as, for instance, position effect. These genetic phenomena reflect, in part, our inadequate knowledge of the structure and properties of proteins and genes, and in part the fact that undoubtedly our present concept of the relationship of gene to enzyme is inadequate. Thus one can say with assurance that the road ahead in the field of biochemical genetics may be different from what we currently envisage, but it is bound to be exciting.

REFERENCES

1. Adelberg, E. A. Bonner, D., and Tatum, E. L., "2, 3-Dihydroxy, 3-Ethyl-*n*-butyric Acid, a Natural Precursor of Isoleucine," in press.
2. Avery, O. T., McLeod, C. M., and McCarty, M., *J. Exceptl. Med.*, **79**, 137 (1944).
3. Beadle, G. W., *Chem. Revs.*, **37**, 15 (1945).
4. ———, and Tatum, E. L., *Proc. Nat. Acad. Sci.*, **27**, 499 (1941).
5. ———, Mitchell, H. K., and Nyc, J. F., *Proc. Nat. Acad. Sci.*, **33**, 155 (1947).
6. Boivin, A., *Cold Spring Harbor Symposia Quant. Biol.*, **12**, 6 (1947).
7. Bonner, D., *Am. J. Botany*, **33**, 788 (1946).
8. ———, *J. Biol. Chem.*, **166**, 545 (1946).
9. ———, *Proc. Nat. Acad. Sci.*, **34**, 5 (1948).
10. ———, *Science*, **108**, 735 (1948).
11. ———, and Beadle, G. W., *Arch. Biochem.*, **11**, 319 (1946).

12. ———, and Yanofsky, C., *Proc. Nat. Acad. Sci.*, **35,** 576 (1949).
13. ———, Tatum, E. L., and Beadle, G. W., *Arch. Biochem.*, **3,** 71 (1943).
14. Burkholder, P. R., and Giles, N. H., *Am. J. Botany,* **34,** 345 (1947).
15. Caspersson, T., *Soc. Exceptl. Biol. Symposium,* **1,** 127 (1947).
16. Dodge, B. O., *J. Agr. Research,* **35,** 289 (1927).
17. Fries, N., *Physiologia Plantarum,* **1,** 330 (1948).
18. Garrod, A. E., "Inborn Errors of Metabolism," 2nd ed., 216 pp., London, Oxford Univ. Press, 1923.
19. Henderson, L. M., *J. Biol. Chem.,* **178,** 1005 (1949).
20. Horowitz, N. H., *Proc. Nat. Acad. Sci.,* **31,** 153 (1945).
21. ———, *J. Biol. Chem.,* **162,** 413 (1946).
22. ———, *J. Biol. Chem.,* **171,** 255 (1947).
23. ———, Bonner, D., and Houlahan, M. B., *J. Biol. Chem.,* **159,** 145 (1945).
24. ———, ———, Mitchell, H. K., Tatum, E. L., and Beadle, G. W., *Am. Naturalist,* **79,** 304 (1945).
25. Knight, B. C. J. G., "Bacterial Nutrition" Med. Research Council (Brit.), Special Report Series No. 210, London, 1936.
26. Krehl, W. A., Teply, L. J., Sarma, P. S., and Elvehjem, C. A., *Science,* **101,** 489 (1945).
27. Lederberg, J., *Genetics,* **32,** 505 (1947).
28. ———, *Records Genetics Soc. Am.,* **17,** 46 (1948).
29. ———, and Tatum, E. L., *Nature,* **158,** 558 (1946).
30. Lindegren, C. C., *Iowa State College J. Sci.,* **16,** 271 (1942).
31. Lwoff, A., *Ann. Inst. Pasteur,* **61,** 580 (1938).
32. McCarty, M., *Bact. Revs.,* **10,** 63 (1946).
33. Mirsky, A. E., and Ris., H., *J. Gen. Physiol.,* **31,** 7 (1947).
34. Mitchell, H. K., and Lein, J., *J. Biol. Chem.,* **175,** 481 (1948).
35. ———, and Nyc, J. F., *Proc. Nat. Acad. Sci.,* **34,** 1 (1948).
36. Monod, J., *Growth Symposium,* **11,** 223 (1947).
37. Oparin, A. I., "The Origin of Life," trans. S. Morgulis, 270 pp., New York, Macmillan Co., 1938; see also Alexander, J., "Life—Its Nature and Origin," New York, Reinhold Publishing Corp., 1948.
38. Perkins, D., Thesis, Columbia University, 1949.
39. Pomper, S., Thesis, Yale University, 1949.
40. Schopfer, W. H., "Plants and Vitamins," Waltham, Mass., Chronica Botanica Co., 1943.
41. Singal, S. A., Briggs, A. P., Sydenstricker, V. P., and Littlejohn, J. M., *J. Biol. Chem.,* **166,** 573 (1946).
42. Spiegelman, S., *Cold Spring Harbor Symposia Quant. Biol.,* **11,** 256 (1946).
43. Srb, A. M., and Horowitz, N. H., *J. Biol. Chem.,* **154,** 129 (1944).
44. Stanley, W. M., "Science in Progress," Ser. VI, Chap. VII, New Haven, Yale Univ. Press, 1949.
45. Tatum, E. L., and Bonner, D., *Proc. Nat. Acad. Sci.,* **30,** 30 (1944).
46. ———, and Beadle, G. W., *Ann Misssouri Botan. Garden,* **32,** 125 (1945).
47. Umbreit, W. W., Wood, W. A., and Gunsalus, I. C., *J. Biol. Chem.,* **165,** 731 (1946).
48. Wagner, R. P., *Proc. Nat. Acad. Sci.,* **35,** 185 (1949).
49. ———, and Guirard, B. M., *Proc. Nat. Acad. Sci.,* **34,** 398 (1948).
50. Wright, Sewall, *Physiol. Revs.,* **21** (3), 487 (1941).
51. Yanofsky, C., and Bonner, D., *Proc. Nat. Acad. Sci.,* **36** (1950).

CHEMICAL PROBLEMS IN THE GENETICS
OF DISEASE PROCESSES *

John W. Gowen

Genetics Dept., Iowa State College, Ames, Iowa

DISEASE, from the modern viewpoint, may be regarded as the resultant of three factors: (1) the constitution of the host, (2) the disease-producing entity, the pathogen, and (3) the environment, which, in a broad sense, may favor either the host or pathogen. In this complex, the geneticist faces the same problems of reacting systems that the chemist faces, possibly on a lesser scale, in his every-day work. The units are on a larger scale. Instead of atoms, the genes probably are represented by large protein molecules. They are no less fixed entities, however. They change spontaneously or mutate. Molecules of large size likewise change to other forms and gain new properties. They are distinct from one another, guiding chemical reactions into distinct channels. They work on substrates. Again the problem is how they lead to the different reacting systems.

In higher forms, the normal host is the resultant of the action of some thousands of genes (5,000 to 30,000 with possibly 8,000 to 10,000 as current estimates in *Drosophila*—Gowen and Gay, 1933). The normal host is brought into being by the mutual interaction of these genes on substrates of familiar substances, fat, carbohydrate, protein, etc. The species form recognized at birth is brought into being through the fetal development of a seemingly undifferentiated egg.

It is obvious that each species is guided to its type by its complex of characteristic genes. It is not so obvious, but probably true, that each individual in the species is likewise distinct from every other, through the action of a small number of genes within the greater species complex, or by the fact that no gene system is exposed to the identical environmental conditions.

When a gene mutates or environmental effects become extreme, morbid development may result. The morbid condition may be due to a gene or a group of genes substituted for the normal genes—their alleles—which work on a tissue substrate formed by the remaining normal genes. In this substitution, the exchanged gene is, in truth, a pathogen as it parasitizes the normal development of the host and leads to its death or unfitness for its struggle for survival against the hosts having the normal alleles. The severity of these effects depends on the surroundings. Favorable environments may alleviate the effects of a bad inheritance of semi-lethal genes. Poor environments may increase the severity of gene effects to a point where death ensues. In a complex of 10,000 normal genes, one causing abnormal growth or physiological effects may lead to death.

The morbid condition may require another factor, a pathogen—virus, bacterium or worm—acting on the host to bring about a disease. The pathogen, like the host, has its actions controlled by its genes. These genes generate the structures necessary for the different apparatus used by the variety of pathogens which attack the same or different hosts. An infectious disease syndrome rests on three legs: the host's gene complexes which favor resistance or susceptibility, the pathogen's gene attributes which lead to differences in virulence, and the environmental complexes which may increase or reduce the severity of the different host-pathogen relations.

Host Responses to Inferior Genetic Constitutions

An early example of the part played by the host's genetic constitution in pathological conditions was found by Cuénot [2] in the yellow mouse coat color. The yellow mice differed from most of the others in the fact that when they were bred together, some of the progeny always showed nonyellow coat colors. The ratio of yellow to nonyellow in the progeny of two yellow mice was 2:1 instead of the expected 3:1. Again, if yellow mice were bred to nonyellow, progeny having either yellow or nonyellow coat colors were obtained in nearly equal proportions. This is the proper proportion expected if all yellow mice have the genes Yy instead of YY. But what became of the YY progeny? Investigations of the uteri of the yellow mice bred to yellow males showed that the death of the YY embryos occurred early in embryological development. The Y gene, working on the rest of the complex of mouse genes in heterozygous condition, Yy, caused the yellow color. When homozygous, YY, it caused development so abnormal as to lead to death. The hereditary endowment given to an embryo by its parents, even to that contributed by a single pair of genes, was capable of ending the life it had a part in beginning.

But not all genes, like the yellow gene in homozygous condition, YY, lead to death before birth. The yellow gene, when in heterozygous state, Yy, gives a fully viable yellow mouse. At birth, the Yy and yy mice are closely similar in appearance save that the former have yellow hair. This similarity continues to nearly 50 days from birth. From 50 days onward, the yellow mice become progressively fat until at 200 days they weigh around 50 grams while their nonyellow litter mates, yy, weigh 30 grams. This difference in weight is accounted for almost entirely by fat deposition and its accompanying water. It is accomplished by only a slight increase in food consumption and reduction in body activity (Dickerson and Gowen,[3]). The increase in fat is paralleled by increasing sterility, particularly in the females.

Human data are noted for cases where the heterozygous condition of such pathological genes seriously handicaps the individual or leads it to its death. Macklin[4] cites 3500 references and 210 lethal or semi-lethal conditions attributable to simple inheritance. In *Drosophila*, more than 10,000 pathological changes attributable to lethal or semi-lethal inheritance have been studied. The number of these cases found in the species is apparently limited only by the study put upon them. An almost unlimited variety of pathological conditions is the result of one to a limited number of genes found among the otherwise normal complex in the fertilized egg. It is well to recall that these lethal or semi-lethal genes are the alleles of genes leading to normal development. Table 1 describes a few of the effects of the presence of semi-lethal genes in otherwise normal individuals of different species.

The problem for the geneticist, as for the chemist, is that of tracing the series of reactions by which these results are accomplished. These reactions are expressed in histopathological changes. Real progress has been made in clarifying the steps by which the end phenotype is reached. A dwarf condition in mice is a case in point. The dwarf mice weigh 8 to 14 grams when their litter mates weigh 25 to 40 grams. They are sterile in both sexes and show thyroid, adrenal and reproductive repression. Basal metabolism is reduced to one-half normal. Transplantation of pituitaries of normal mice into the dwarfs caused a resumption in growth and a return of the thyroids, adrenals and reproductive tract to normal. The males became fertile but the females remained sterile. The gene had acted through its effect on the pituitary function. Extracts from normal pituitaries of other species were likewise able to bring the dwarf mice to normal size and function. The replacement therapy had supplemented the defective inheritance and brought the animal into nearly normal function, but it had in no way influenced the gene or its primary effect. The eosinophil cells of the pituitary were still only a small fraction of their normal number. So, too, the progeny of the treated males were

TABLE 1. GENETIC CONDITIONS ATTRIBUTED TO SINGLE INHERITANCE DIFFERENCES, SINGLE GENES

Condition	Inheritance	Animal	Investigator
Agnathia	r	Cattle	Annett[5]
Fused teats	r	Cattle	Johnson[6]
Congenital lethal spasms	r	Cattle	Gregory, Mead, and Regan[7]
Polydactylism	D	Cattle	Roberts[8]
Epithelial defects	r	Cattle	Hutt, and Frost[9]
Nakedness	D	Chicken	Sturkie[10]
Brachydactyly	r	Chicken	Warren[11]
Dactylolysis	r	Chicken	Shoffner[12]
Short mandible	r	Dog	Gruneberg, and Lea[13]
Hemophilia	r	Dog	Hutt, Rickard, and Field[14]
Paralysis—lethal	D	Dog	Stockard[15]
Congenital palsy	r	Guinea pig	Cole, and Ibsen[16]
Transplacental isoimmunization	D ?	Horse	Levine[17]
Huntington's chorea	D	Human	Davenport[18]
Xeroderma pigmentosum	r, I	Human	Dresel[19]
Rh-blood factor	D	Human	Landsteiner, and Wiener[20]
Alkaptonuria	r	Human	Hogben, Worrall, and Zieve[21]
Chrondrodystrophia foetalis	D	Human	Morch[22]
Polydactylism	r ?	Human	Snyder[23]
Phenylketonuria	r	Human	Penrose[24]
Flexed tail	r	Mouse	Mixner, and Hunt[25]
Short tail—urogenital defect	D	Mouse	Danforth[26]
Lens rupture	r	Mouse	Fraser, and Herer[27]
Amputated	r	Pig	Johnson[28]
Hemophilia-like	r	Pig	Bogart, and Muhrer[29]
Syndactylism—mule foot	D	Pig	Detlefson, and Carmichael[30]
Porphyrinuria	r	Pig	Clare, and Stephens[31]
Polydactyly of feet and wings	r	Pigeon	Hollander, and Levi[32]
Ataxia	r	Pigeon	Riddle, and Hollander[33]
Microphthalmia (small eye)	r	Pigeon	Hollander[34]
Alpha-agglutinins	D	Rabbit	Wheeler, Sawin, and Stuart[35]
H₁ H₂ Blood groups	D	Rabbit	Castle, and Keeler[36]
Epilepsy	r	Rabbit	Nachtsheim[37]
Prognothism	D	Rabbit	Nachtsheim[38]
Shaking paralysis	r	Rabbit	Nachtsheim[39]
Anemia	r	Rat	Smith, and Bogart[40]
Cartilage lethal	r	Rat	Gruneberg[41]
Wobbly	r	Rat	Castle, King and Daniels[42]
Photosensitivity	r	Sheep	Hancock, and Basset[43]
Yellow fat	r	Sheep	Castle[44]
Nervous incoordination	r	Sheep	Rasmussen[45]

D—Dominant
r—Recessive
I—Incompletely dominant

dwarfs of the same type as their parents, having received the gene for the condition unchanged from that transmitted to their fathers. As has always proven so in adequately analyzed cases, there is no transfer of the acquired growth recovery to the untreated offspring. The gene remains unmodified.

Alkaptonuria, another inherited condition, has had its effects identified with recognized chemicals. The Alkaptonuric differs from the normal man in being homozygous for a recessive gene pair. People having this

inheritance lack a specific enzyme and cannot break down the 2,5-dihydroxyphenyl acetic acid of normal catabolism. By oxidation, this substance appears in the urine as dark pigment. The function of the normal gene appears to be that of forming the enzyme to complete this chemical breakdown.

The cases cited in Table 1, will help to recall the multitude of inherited effects which may be traced to gene differences within the host. For any one gene, the syndrome of effects produced is fairly specific, but the corollary does not follow that a pathological condition is, in itself, diagnostic of the presence of a particular gene. Haldane [46] illustrates that fact with an apt case. He points out that one of the common causes of blindness is retinitis pigmentosa. This condition may be due to an autosomal dominant gene in one case, a recessive autosomal gene in a second, a recessive sex-linked gene in a third, and a partial sex-linked gene in a fourth. The blindness in each case is similar, yet there is evidence in other effects such as those on hearing, that the genes control different stages in the development of sight. Pathologists will need to work out the different types, but it is certain that this cannot be done unil the gene responsible is known, for it is probable that this process through which each reaches its common goal, blindness, is different and will require different treatment for its cure. Symptoms are no more a criterion of identity of causes in genetic diseases than they are in dysenteries of as different etiology as bacillary and amoebic or in such common characteristics as color of eyes in *Drosophila*. Vermilion eyes are converted to wild type brick-red by an injection of kynurenine to replace the lost oxidation process of tryptophan to kynurenine normally occurring in vivo. Cinnabar is not so cured, as this gene affects a later stage in pigment formation. Scarlet and cardinal are not cured because they lack pigment precursors. The four genes all affect eye pigment and each is located in a different place in the chromosome complex. They have a certain similarity of symptoms but their inheritances and their chemical actions on the eye-forming process are different. As Gowen pointed out some time ago, the facts in diseases of pathogenic origin are in the same category. Although the disease syndrome is similar by the actual reaction, systems may be very different from individual to individual, owing to differences in genotypes of both the hosts and the strain of the pathogens. These differences may require radically different treatments to restore the individuals to health.

Work on chemical genetics by Onslow, Scott-Moncrieff, Ephrussi, Beadle, Tatum and others is clearing up some of these problems. Not all disease effects are traced to host genes. Many disease reactions turn on inherited differences in a pathogenic species or in the interaction of

the host and pathogen genotypes. The chemistry of preventive and curative drugs is helping to clarify the problem.

Host Response to Infectious Disease

In animals, instances where susceptibility or resistance to a disease is dependent on a single pair of genes are rare; yet some would be expected to occur, as the pathogen as well as the host has its characteristics prescribed by its ancestors. An epidemic disease of mice is a possible example. The pathogen implicated is a long, slender bacillus, *B. piliformis.* This pathogen has strict growth preferences for the livers of particular mice. In nature, it attacks certain members of the Asiatic species, *Mus bactrianus.* Our ordinary house mouse, *Mus musculus,* is nearly, if not completely, resistant to the organism. Crosses between the two species lead to F_1's which are resistant. Backcrosses to the susceptible Asiatic species give progeny which segregate for resistance and susceptibility in nearly the right proportions to account for the differences in resistance between the species as due to a single gene pair with resistance dominant. The closely established growth requirements of the pathogen suggest that the genes responsible may affect a single chemical bond in the metabolic processes within the animal, leading to the production of a chemical required in the growth of *B. piliformis.*

A like case is found in guinea pigs. In routine Wassermann tests in Vermont, Rich [47] noted guinea pigs which had no blood complement. Breeding tests showed that the characters, full and deficient complement, depended on a single gene pair, the deficient complement being recessive. A spontanous epidemic of *B. suisepticus* within this mixed colony presented an opportune test for the part played by complement in resistance to this disease. Full-complement guinea pigs often developed lesions of the liver, spleen and peritoneum, but seldom died. The deficient-complement pigs, on the other hand, died so suddenly that they did not even develop characteristic lesions. The focus of the epidemic was the deficient-complement pigs. An experiment to test this matter was performed. One hundred pigs of each group were inoculated with living *B. cholerae suis* under similar conditions. Seventy-seven of the 100 complement-deficient pigs were dead in 48 hours; twenty of the full-complement pigs died in the same period. The two experiments clearly show the significance of a single pair of genes to survival of their hosts when exposed to this disease. The effect was produced through a modification of a single blood-serum chemical, complement. Detection and study of the products of this gene difference furnished significant information on the chemical nature of complement. Full-complement pigs were either homozygous dominants, *Cm Cm,* or heterozygous *Cm cm;* the deficient pigs were homozygous

recessive, *cm cm*. Breeding tests show that mothers whose blood is complement-deficient may have offspring which are normal—have full blood complement—depending on the genotype of the father. Similarly, mothers high in complement *Cm cm* can have offspring which are deficient. The numbers of these offspring meet the requirements of simple Mendelian inheritance. These facts prove that the condition is due to a single gene-pair difference. They show that the product formed, complement, is a molecule of such size that it cannot pass the placental membrane. The data deal a severe blow to the advocates of the inheritance of acquired characters, since the inherited character, the gene for deficient complement, may be associated with the inherited factor, the gene for full complement, for many generations and yet produce its deficient-complement effect unchanged by this long association.

The observed character in the normal guinea pig is complement. The alteration of this character changed the clinical and pathological possibilities of the guinea pigs. Studies by Coca [48] and Hyde [49] have thrown additional light on this change. It is not the whole complement "molecule" which is affected; it is only that part which is heat-stable at 56°C for 30 minutes. Tests by different disease-producing entities gave further information on the functional relations of the groups making up the complement molecule. At first thought, such a genetic constitution as that of the complement-deficient pigs might be considered to have no survival value. This is not the case; Hyde has shown that the complement-deficient guinea pigs are immune to the action of heterophilic sera, sheep, chicken and rabbit, whereas normal-complement pigs die within 2 to 5 minutes after inoculation, with symptoms resembling shock. The gene mutation, from normal complement-forming to complement-deficient, led to two vital effects. The loss of this chemical from the blood resulted in increased susceptibility to certain diseases; the lack of this chemical increased the animal's resistance to a vivo chemical reaction involving heterophilic sera.

As would be expected, many diseases of pathogenic origin show more complex reacting systems. The work on mouse typhoid specifies some of these problems as they occur in the Iowa State College Genetics Laboratory. The starting material was a single strain of the disease-producing organism, *Salmonella typhimurium*, the pathogen of mouse typhoid and a heterogeneous group of mice of no particular breeding. By selecting survivors within one mouse group, it was possible to establish in fifteen generations a strain which was 95 to 98 per cent resistant. Other strains, which were established through inbreeding alone, were found to have characteristic resistances ranging from 5 to 8 to 75 to 85 per cent. These differences in levels of strain resistance remain constant when the strains are exposed to the same number of organisms at the same age and under

the same standard conditions. Inheritance studies show that these strain differences are inherited in a complex manner. The chemistry of this process is unclear, although there are several observations which merit consideration.

Studies of mice in health and in sickness show that the strains differ in disease expression in other ways than percentages of deaths. Bacteria apparently grow rapidly in the susceptible mice, giving rise to a general bacteriemia which usually terminates in death without liver lesions. The large liver lesions of the resistant strains exhibit few or no bacteria, although the effects of highly active toxins are evident. Lesions are separated from normal tissue by a dense concentration of heterophils, macrophages and fibrocytes. This barrier is sufficient to prevent the passage of toxins, etc., from the lesions to the normal cells, as the normal tissue is found to be functioning satisfactorily; glycogen is being formed and stored, fat that is synthesized and deposited takes characteristic dyes. The chemical functions of the liver are maintained for the normal body functions of the rest of the body. Susceptible mice show little gross liver damage but rapid deaths. Glycogen formation ceases in the whole liver. The fat does not stain properly. The liver is evidently in dysfunction, at least for the chemical processes connected with glycogen and fat formation.

The liver and spleen filter out most of the invading typhoid organisms in mice. Strains of mice differing in their typhoid resistance react differently to the presence of the bacteria. The intermediate and highly resistant strains show severe liver lesions and less pronounced splenic lesions. Susceptible strains have few, if any, liver lesions. The macrophages in the spleens of different strains react differently to the presence of the bacteria. Large numbers of bacteria are phagocytosed by the splenic macrophages of the susceptible mice. These bacteria stain and otherwise appear normal within the macrophage cells. Single cells may contain 30 to 50 apparently healthy, reproducing bacteria. In fact, it appears as if the bacteria might reproduce within the cell to such an extent as to destroy the cell and release the bacteria to further the disease process. The macrophages of the resistant strains of mice react quite differently. Bacteria in these macrophages are difficult to observe. When found, these bacteria do not stain properly and the cell outlines appear ragged. The bacteria look as if they were undergoing rapid digestion by an intracellular enzyme. The enzyme which the observations suggest, could be generated in the resistant mice by the specific gene inheritance they receive, whereas the susceptible mice lacking the proper inheritance would lack the enzyme. This view is in agreement with modern views on the chemical action of genes in general. It raises problems of isolation,

demonstration, and structure of the hypothetical enzyme, together with the mechanism of its action.

Similar chemicogenetic problems enter into acquired resistance to a disease gained through vaccination. The same three elements, host constitution, pathogen constitution, and environment contribute to the initiation and severity of a disease following vaccination as were important to the animal prior to it. Vaccinating our six strains of mice, differentiated for normal typhoid resistance, with three lines of *Salmonella typhimurium* ranging from nearly saprophytic to highly virulent strains, gave the following results after vaccination. Vaccinated host strain differences, following a massive challenge dose of virulent live organisms, show direct parallelism to the genetic resistance of the strains on first exposure to the disease. Physiologically, the genetically controlled elements responsible for natural resistance appear to be those which are likewise important to the acquisition of resistance through vaccination. These observations materially affect the search for the systems involved. Genotype of the pathogen has a comparable effect within each mouse strain. The genetic elements leading to virulence or avirulence form one part of the reacting system; the other part is the constitutions of the hosts. The interactions of these systems give health, morbidity, or death, according to the particular position occupied within the system by the given host-pathogen genetic complex.

REFERENCES

1. Gowen, J. W., and Gay, E. H., *Genetics*, 18, 1-31 (1933).
2. Cuénot, L., *Arch. zool. exptl. et gén.*, 3, 123-32 (1905).
3. Dickerson, G. E., and Gowen, J. W., *Science*, 105, 496-8 (1947).
4. Macklin, M. T., *Medicine*, 14, 1-75 (1935).
5. Annett, H. E., *J. Genetics*, 37, 301-2 (1939).
6. Johnson, L. E., *J. Heredity*, 36, 317-20 (1945).
7. Gregory, P. W., Mead, S. W., and Regan, W. M., *J. Heredity*, 35, 195-200 (1944).
8. Roberts, Elmer, *J. Heredity*, 12, 84-6 (1921).
9. Hutt, F. B., and Frost, J. N., *J. Heredity*, 39, 131-7 (1948).
10. Sturkie, P. W., *J. Heredity*, 33, 202-8 (1942).
11. Warren, D. C., *J. Heredity*, 31, 141-4 (1940).
12. Shoffner, R. N., *J. Heredity*, 36, 375-8 (1945).
13. Gruneberg, H., and Lea, A. J., *J. Genetics*, 39, 285-96 (1940).
14. Hutt, F. B., Rickard, C. G., and Field, R. A., *J. Heredity*, 39, 3-9 (1948).
15. Stockard, C. R., *Am. J. Anat.*, 59, 1-53 (1936).
16. Cole, L. J., and Ibsen, H. L., *Am. Naturalist*, 54, 130-51 (1920).
17. Levine, P., *J. Heredity*, 39, 285-8 (1948).
18. Davenport, C. B., *Proc. Nat. Acad. Sci.*, 1, 283-5 (1915).
19. Dresel, K., *Virchow's Arch. Path. Anat.*, 224, 256-303 (1917).
20. Landsteiner, K., and Wiener, A. S., *Proc. Soc. Exptl. Biol. Med.*, 43, 223-4 (1940).
21. Hogben, L., Worrall, R. L., and Zieve, I., *Proc. Roy. Soc. Edinburgh*, 52, 264-95 (1932).
22. Morch, E. T., *J. Heredity*, 31, 439-44 (1940).
23. Snyder, L. H., *J. Heredity*, 20, 73-7 (1929).

24. Penrose, L. S., *Lancet,* 229, 192-4 (1935).
25. Mixner, R., and Hunt, H. R., *Genetics,* 18, 367-87 (1933).
26. Danforth, C. H., *Proc. Soc. Exptl. Biol. Med.,* 23, 145-7 (1935).
27. Fraser, F. C., and Herer, M. L., *J. Heredity,* 39, 149-50 (1948).
28. Johnson, L. E., *J. Heredity,* 31, 239-42 (1940).
29. Bogart, R., and Muhrer, M. E., *J. Heredity,* 33, 59-64 (1942).
30. Detlefson, J. A., and Carmichael, W. J., *J. Agr. Research,* 20, 595-604 (1921).
31. Clare, N. T., and Stephens, E. H., *Nature,* 153, 252-3 (1944).
32. Hollander, W. E., and Levi, W. M., *J. Heredity,* 33, 385-91 (1942).
33. Riddle, O., and Hollander, W. F., *J. Heredity,* 34, 167-72 (1943).
34. Hollander, W. F., *J. Heredity,* 39, 289-92 (1948).
35. Wheeler, K. M., Sawin, P. B., and Stuart, C. A., *J. Immunol.,* 36, 349-59 (1939).
36. Castle, W. E., and Keeler, C. E., *Proc. Nat. Acad. Sci.,* 19, 92-8 (1933).
37. Nachtsheim, H., *Z. menschl. Vererbungs- u. Konstitutionslehre,* 22, 791-810 (1939).
38. ———, *Deut. tierärztl. Wochschr.,* 44, 742-6 (1936).
39. ———, *Erbarzt (Beil. z. deut. Aerztebl.)* 1, 36-8 (1934).
40. Smith, S. E., and Bogart, R., *Genetics,* 24, 474 (1939).
41. Gruneberg, H., *Proc. Roy. Soc. London,* 125B, 123-44 (1938).
42. Castle, W. E., King, H. D., and Daniels, A. L., *Proc. Nat. Acad. Sci.,* 27, 250-4 (1941).
43. Hancock, J., and Bassett, E. G., *Proc. New Zealand Soc. An. Prod.,* 6, 90-5 (1946).
44. Castle, W. E., *J. Heredity,* 25, 246-7 (1934).
45. Rasmussen, K., *Sci. Agr. (Ottawa),* 25, 482-8 (1945).
46. Haldane, J. B. S., *Proc. Royal Soc. (London),* 135B, 147-70 (1948).
47. Rich, F. A., *Vermont Agr. Expt. Sta., Bull.,* 230, 24 (1923).
48. Coca, A. F., *Proc. Soc. Exptl. Biol. Med.,* 18, 71 (1920).
49. Hyde, R. R., *J. Immunol.,* 8, 267-86 (1923).

ANTIBIOTIC COLLOIDAL ELECTROLYTES

Ernst A. Hauser

*Professor of Colloid Chemistry, Mass. Institute of Technology, Cambridge, Mass.
and Worcester Polytechnic Institute, Worcester, Mass.*

Introduction

In 1912, McBain and his collaborators reported that a pure salt of a higher fatty acid—a soap—behaves like an ordinary electrolyte in very dilute solution, but exhibits properties typical for colloidal sols with increasing concentration.[15] To distinguish them from true electrolytes, McBain termed such substances *colloidal electrolytes* and defined them as electrolytes in which one ion is replaced by conducting colloidal particles which are spontaneously formed in the solution.[16-18] When he presented some of the evidence for the formation of such colloidal electrolytes to the Colloid Committee of the British Association for the Advancement of Science, in London, in 1925, it was dismissed by the Chairman with the word, "nonsense." [16]

In 1947, it was stated for the first time by Hauser and his collaborators that solutions of certain antibiotics, like penicillin and streptomycin, must also be classified as colloidal sols and not as true solutions.[9] Soon, thereafter, the results of more systematic research into the properties exhibited by solutions of antibiotics were submitted to gatherings of colloid chemists.[10, 11] During the discussions immediately following the presentation of these papers, and since then in scientific literature, these new findings have been subjected to severe criticism.[13, 14, 19, 23]

Since the problem of how antibiotics act is of more than academic interest, and their classification as colloidal electrolytes might make a considerable difference in explaining their activity, the following facts are offered as proof for their colloidality. Besides this it will once again prove how correct Darwin (1809-1882) was in his statement, "False facts are highly injurious to the progress of science; for they often endure long; but false views, even if supported by some evidence, do little harm, for everyone takes a salutary pleasure in proving their falseness." [1] It is

therefore important to explain why the criticism voiced against classifying antibiotics as colloidal electrolytes is based on false facts.

Chemical and Physical Aspects

The basic formula for the most important penicillin preparations now in use is shown in Figure 1. The only significant difference from other suggested formulas lies in the composition of the hydrocarbon chains represented by R (Table 1) and the possible exchange of the H$^+$ ion of

$$
\begin{array}{c}
\text{S} \\
\diagup\diagdown \\
\text{O} = \text{C} - \text{HN} - \text{CH} - \text{CH} \qquad \text{C (CH}_3)_2 \\
\mid \qquad\qquad \mid \qquad \mid \qquad\qquad\qquad \mid \\
R \qquad\quad \text{O} = \text{C} --- \text{N} ------ \text{CH} - \text{COOH}
\end{array}
$$
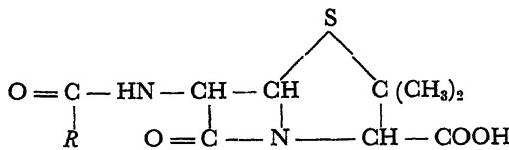
Figure 1. Basic formula for penicillic acid. R is hydrocarbon chain.

the hydroxyl group of the acid radical with other cations, resulting in the formation of salts, like sodium- or ·calcium penicillin. Molecular-weight determinations have been carried out by determining the diffusion constant [12] and by measuring freezing-point depressions. The molecular weight obtained ranged between 312 and 490, depending on the origin of the sample. Conductance measurements and osmotic activities of sodium penicillins in aqueous solutions of low concentrations have also been reported.[10-12, 17, 19, 22] On the basis of all these data it has been

TABLE 1.

Penicillin	R
F	CH_3 CH_2 $CH{=}CH{-}CH_2{-}$
G	C_6H_5 $CH_2{-}$
X	$p{-}HOC_6H_4$ $CH_2{-}$
Dihydro-F	CH_3 CH_2 CH_2 CH_2 $CH_2{-}$
K	CH_3 CH_2 CH_2 CH_2 CH_2 CH_2 $CH_2{-}$

concluded that penicillin salts in solution must be classified as simple, or completely dissociated, electrolytes.[12-14, 19, 23,26] Although some deviations owing to ion size and interactions were taken into consideration, the results were in general found to be in line with those to be expected from a 1:1 valence-type electrolyte. Similar results have also been reported for streptomycin and other antibiotics.

Colloid Chemical Aspects

If the above findings are compared with those made with substances termed micelle- or association colloids, or colloidal electrolytes today,[16]

their similarity becomes evident. As stated in the introduction, the salts of the higher fatty acids, the soaps, are typical colloidal electrolytes. In extreme dilutions only highly disperse systems are formed, but with increasing concentration the degree of ionization drops and the particle size of the micelle increases, owing to association of molecules, until it reaches colloidal dimensions. Decreasing temperature aids, increasing temperature diminishes, micelle formation.

While very dilute solutions of penicillin or streptomycin salts show only a bluish light cone if studied at room temperature in the ultramicroscope, concentrations as used for medical treatment reveal a clear Faraday-Tyndall cone. If the temperature of the solution is dropped below 20°C, even the very dilute solutions give rise to a Faraday-Tyndall effect caused by the formation of many colloidal particles of anisometric shape.
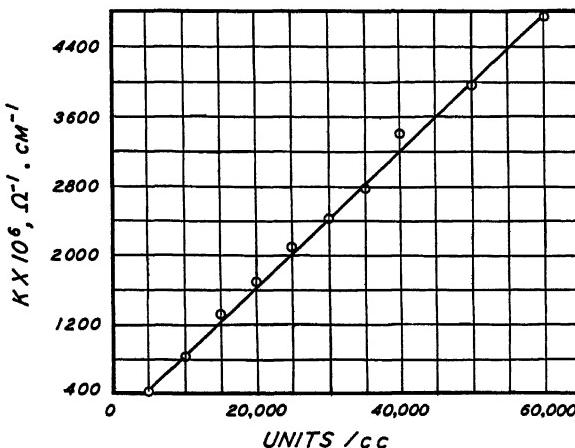
Figure 2. Conductivity of sodium penicillin at various concentrations.[3]

The addition of minute quantities of electrolyte with multivalent cations, as for example, $COCl_3$, $AlCl_3$, or $ThCl_4$, to solutions of crystalline sodium penicillin increases immensely the number of ultramicroscopically-detectable colloidal particles. If such a colloidal sol is subjected to ultracentrifugation, the system separates into two layers. One layer shows a high concentration of anisometric particles, whereas the other is optically void of colloidal matter. Microanalysis of these fractions reveals that the amount of sodium found in the continuous phase is exactly in the 3:1 ratio for the amount of Al^{+++} or Co^{+++}, and in a 4:1 ratio for the Th^{++++} ions which had been added. This clearly indicates that an ion exchange reaction has taken place. The greater number of micelles

formed is attributed to the lesser hydration of the exchanging ions and the less pronounced ionization resulting therefrom.

Specific conductivity tests of highly purified sodium penicillin solutions show a linear rise with increasing concentrations (Figure 2).[10] The penicillin salt is a colloidal electrolyte composed of ions, one of them being so large that it will not dialyze; but the number of ions and their charge still cause measurable conductivity. The condition therefore is comparable to that typical for normal electrolyte solutions.[5] The same results have also been found with streptomycin.
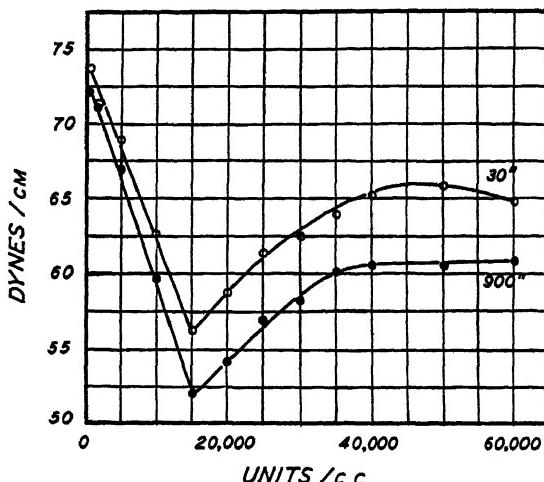
Figure 3. Surface tension of crystalline sodium penicillin at various concentrations and times, determined with the pendant-drop method at 37°C.

Surface-tension determinations of solutions of sodium penicillin carried out with the pendant-drop method not only revealed that this is a capillary-active substance, but that it must be classified as a colloidal electrolyte because its surface tension vs. concentration curve, in contrast to that of a normal capillary-active substance, shows a minimum [10] characteristic for colloidal electrolytes (Figure 3).[17] Streptomycin complex salts give similar, although not so pronounced, effects.

Another property which permits differentiating between true electrolytes, colloidal electrolytes composed of anisometric particles, and macromolecular colloids, is the relationship between viscosity and concentration. Whereas true electrolytes show practically no increase in viscosity with increasing concentration, and macromolecular colloids show a proportional increase, the viscosity of colloidal electrolytes increases far less.[22] Viscosity determinations of sodium penicillin sols gave

the same results as those reported for other colloidal electrolytes (Figure 4).[10]

A few years ago some work was carried out to ascertain whether electrokinetic potential studies of bacteria to which bacteriostatic agents had been added could shed some light on the question of what causes their activity.[2] Although a pronounced change in the potential of the bacteria was noticed upon addition of the agent, the conclusion was drawn that this could not have been the result of adsorption of the bacteriostatic agent on the surface of the bacterium. This opinion was based on the observation that the electrokinetic potential of quartz particles added to a similar medium was not affected.
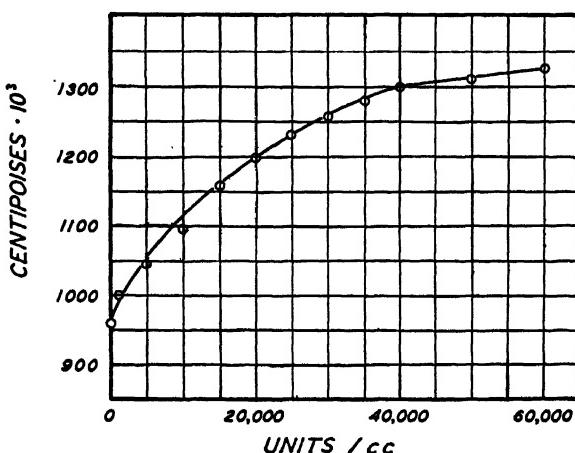
Figure 4. Viscosity versus concentration curve of sodium penicillin.

More recently, medical circles reported that penicillin causes a change of the electrostatic charge of susceptible bacteria which is measurable in terms of their zeta potential. It has also been claimed that bacterial growth in penicillin assay plates showed enhancement in electronegative and inhibition in electropositive zones.[3, 25] These findings add further proof to the fact known to colloid chemists that bacteria act as cation exchangers. The antibiotic action of penicillin might therefore be comparable to that reported for surface-active cations. It has also been ascertained that the addition of 2 to 6000 units of penicillin/ml changes a normal *Staphylococcus aureus* from being a Gram-positive to a Gram-negative organism.[6]

That the antibiotic is actually adsorbed on the surface of the bacterial cell has recently been demonstrated by the use of an ultraviolet-microscope. Penicillin salts fluoresce if exposed to ultraviolet radiation; if a nonfluorescing bacterial culture to which penicillin has been added is

studied with an ultraviolet-microscope, the individual cells will show a fluorescing surface, while the intensity of fluorescence in the dispersion medium decreases.[10, 11]

Recently, several other investigators have reported surface-tension measurements on sodium penicillin solutions containing 10,000 units/cc or 0.017 mole/liter. Lund and collaborator [14] have this to say: "In 1948 Kumler and Alpen, employing both du Noüy's precision tensiometer and the capillary-rise method, carried out surface-tension measurements on aqueous solutions of crystalline sodium penicillin G and crystalline potassium penicillin G and found that solutions of penicillin G have a surface tension differing only little from that of water. Therefore, the solutions must be true solutions and not colloidal sols."

For sodium penicillin G at a concentration of 0.6 per cent and a temperature of 23°C, Kumler and Alpen [13] obtained a surface tension of 70.8 dynes/cm with the du Noüy tensiometer, and of 70.3 with the capillary-rise method. McBain and collaborators [19] also carried out surface-tension determinations of sodium penicillin G, but did not refer to the technique they used, nor did they give any results of their surface-tension measurements. They only state that above a concentration of 0.25 mole/liter penicillin must be classified as a colloidal electrolyte; in lower concentrations, as an ordinary electrolyte. They attribute the fairly low surface tension they found, which was relatively independent of concentration, to small amounts of impurities.

The results obtained with the pendant-drop method [9-11] cannot be attributed to impurities, however, because all measurements were carried out with noncontaminated, absolutely fresh solutions made from the purest salts now obtainable. What seemingly has been overlooked is the fact that when measuring the surface tension of capillary-active substances in solutions of low concentration, the time factor must also be taken into consideration. This was noticed by du Noüy,[20, 21] who had found that certain colloidal substances, even at very high dilutions, lowered the surface tension of water substantially, as a function of time; or, in other words, that the static value was lower than the dynamic value. Later this was substantiated by McBain and his collaborators by the use of the PLAWM trough.[17] Du Noüy must also be given credit, however, for having been the first to point out the shortcomings of his own tensiometer when applying it to the determination of surface tension of very dilute colloidal solutions. Discussing the causes of error in his technique, he states: ". . . The surface of the solution was disturbed every time a measurement was made, and consequently the successive measurements did not exactly express what the value of the surface tension would have been had the surface not been disturbed by the preceding measurement. In making a measurement, the surface of the liquid is so much

distorted that it was by no means certain whether this did not change the state of the surface layer and render the condition of rupture quite different from that which would have been obtained on an undisturbed surface. The distortion may have increased the distance between molecules, and this so rapidly that there was no time to reach an equilibrium before the film was torn." [20] This was later substantiated and visually demonstrated by Hauser and his collaborators, who applied a high-speed motion-picture camera to record their observations.[7] Du Noüy also made several suggestions as to how one could minimize this obvious shortcoming of his technique. Unfortunately, no, or very little, attention has been paid to them by those who use the du Noüy tensiometer for the determination of the surface tension of dilute solutions of capillary-active substances. Of all methods known and in use today for the determination of surface- and interfacial tension, the pendant-drop method [8] is the only one which permits measurements to be made without disrupting or distorting the surface. Besides this, it is the only method which enables one to follow visually any change in the surface tension of a solution with time, and is therefore the only truly reliable one when dealing with ionizable, capillary-active substances in solutions of low concentration. That determinations of osmotic activities,[14] freezing point depression,[13] and even of equivalent conductivity [19] give results in line with those obtained when studying true solutions, may not be considered sufficient evidence to claim that solutions of penicillin salts are true solutions. These results must be considered in conjunction with the most accurate surface-tension measurements. A comparison of all data now available clearly indicates that penicillin salts must be classified as colloidal electrolytes.

REFERENCES

1. Darwin, C. R., "The Descent of Man," Chap. 21, London, 1871.
2. Dorfman, W. A., *Nature*, 153, 169 (1944).
3. Dufrenoy, J., and Pratt, R., *Nature*, 160, 819 (1948).
4. Eiseman, B., *J. Exptl. Med.*, 88, 189 (1948).
5. Freundlich, H., "Kapillarchemie," Vol. II, p. 334, Leipzig, Akademische Verlagsgesellschaft, m. b. H., 1932.
6. Gal, E. F., *Bull. Johns Hopkins Hospital*, 83, 154 (1948).
7. Hauser, E. A., Edgerton, H. E., Holt, B. M., and Cox, J. T., Jr., *J. Phys. Chem.*, 40, 973 (1936).
8. ————, Andreas, J. M., and Tucker, W. B., *J. Phys. Chem.*, 42, 1001 (1938); *Ind. Eng. Chem.*, 31, 32 (1939).
9. ————, Phillips, R. G., and Phillips, J. W., *Science*, 106, 616 (1947).
10, ————, ————, ————, and Vavruch, I., *J. Phys. Coll. Chem.*, 53, 287 (1949).
11. ————, *Koll.-Z.*, 111, 103 (1949).
12. Klenow, H., *Acta Chem. Scand.*, 1, 328 (1947).
13. Kumler, W. D., and Alpen, E. L., *Science*, 107, 567 (1948).
14. Lund, C. G., and Pedersen-Bjergaard, K., *Science*, 109, 149 (1949).

15. McBain, J. W., *J. Chem. Soc. (London)*, 101, 2042 (1912); 105, 417, 957 (1914); 113, 825 (1918).
16. ———, in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. V, p. 103, New York, Reinhold Publishing Corp., 1944.
17. ———, Ford, T. F., and Wilson, D. A., *Koll.-Z.*, 78, 1 (1937).
18. ———, in "Advances in Colloid Science," eds. E. O. Kramer *et al.*, Vol. I, p. 99, New York Interscience Publishers, Inc., 1942.
19. ———, Juff, H., and Brady, A. P., *J. Am. Chem. Soc.*, 71, 373 (1949).
20. du Nouy, P. Lecomte, "Colloid Symposium Monographs," Vol. III, p. 25, New York, Chemical Catalog Co., Inc. (Reinhold Publishing Corp.), 1925.
21. ———, "Surface Equilibria of Biological and Organic Colloids" A.C.S. Monograph Series, New York Chemical Catalog Co., Inc. (Reinhold Publishing Corp.), 1926.
22. Philippoff, W., "Viskositaet der Kolloide," Dresden, Th. Steinkopff, 1942.
23. Pratt, R., Private communication.
24. Staudinger, H., "Organische Kolloidchemie," Braunschweig, F. Vieweg und Sohn, 1941.
25. Strait, L. A., Dufrenoy, J., and Pratt, R., *J. Am. Pharm. Assoc.*, 37, 133 (1948).
26. Woodbury, D. T., and Rosenblum, C., *J. Biol. Chem.*, 171, 44 (1947).
Note: Readers specifically interested in the chemistry and structure of penicillin and related compounds are referred to "The Chemistry of Penicillin," ed. H. T. Clarke, J. R. Johnson, and R. Robinson, Princeton, N. J., Princeton Univ. Press, 1949.
A paper dealing with colloidal phenomena in antibiotics was presented by E. A. Hauser and G. J. Marlowe at the 119th meeting of the Am. Chem. Soc., Division of Colloid Chemistry, September, 1949, and will appear in Vol. 54 of the *J. Phys. Coll. Chem.*

THE TUBERCLE BACILLUS *

Léon Grigoraki

Professor of the Faculty of Athens,
Conference Director, École de Hautes Études, Paris

(Translated by Jerome Alexander)

Asexual Phase

THE VARIOUS VEGETATIVE FORMS of the tubercle bacillus, which are called rods, bacilli, giant forms, ramifying, club-like, yeast-like, or granule forms, are all seen in the asexual or active form of this bacillus. It is proposed here to consider the significance, polymorphism, and mode of multiplication of these forms, matters on which investigators have divergent and often contradictory opinions.

Polymorphism. In an earlier publication entitled "Transformisme génétique—Evolution," the author has established certain fundamental principles governing polymorphism and the transmutation of bionts (living things), and these principles are applicable to the interpretation of the morphological variations of the tubercle bacillus. The principles are as follows:

The volume of the primitive forms increases progressively and their surfaces become irregular, with corresponding diminution in their vitality and virulence. These changes are consequent upon structural degradation determined by nutritional differences and the resulting chemical differences. Variations in form and function are merely manifestations of these structural changes.

According to these principles, the tiny, perennial rods are the primitive forms, which under the influence of nutrition undergo structural alteration, turn into larger and often curved bacilli, and finally into giant forms; the vitality and virulence of these forms diminishes in an order and intensity corresponding to the changes in structure and

* Albert Mary contributed a paper to Vol. II of this series (1928), pp. 869-876, entitled "Colloid Chemistry and Tuberculosis." This refers to earlier work on the so-called "virus forms" of *M. tuberculosis*, which are still looked upon askance by those who fail to see that pleomorphism is not uncommon in microorganisms. The present paper outlines later observations and conclusions drawn by its author, who has worked long and extensively in this field.—*Ed.*

form. Mutations can be produced in the course of degradation, by dissociation of partly modified forms, and by their adaptation to another host, thus creating in the parasite a new biological specificity and modifying its descendants.

The bacilli in a single culture undergo their evolutionary changes at different rates, according to their position relative to the milieu; and the changes in form therefore appear at different times, leading to marked polymorphism in the cultures. The terms *polymorphism, pleomorphism,* and *involuted forms,* have the same meaning both morphologically and biologically.

Variations in Staining. The several forms of the tubercle bacillus react differently to stains; the following classes are recognized: hyper acid-resistant; acid-resistant; cyanophiles; basophiles; iodophiles; etc. We do not know the cause of these variations.

Our investigations indicate that the staining capacity of the bacilli is governed by products of lipid nature excreted by the bacilli. These substances are very chromophilic, but vary in location and in physical and microchemical nature. They can be seen in the larger parasites and are excreted by the protoplasm during the early stages of the culture, finally accumulating at the surface of the organisms to form what the author has termed *ectoplasm*. In early stages of the culture, no ectoplasm having accumulated at the surface, dyes penetrate and stain the bacilli uniformly. In later stages, the ectoplasm, accumulated at the surface, forms a layer which, because of its chromophilic nature, seizes the dye, so that now only the surface is stained in a pronounced and selective manner.

In media encouraging parasitic growth (which contain substances of lipid nature) ectoplasm is abundant; whereas in media encouraging saprophytic growth (which have little lipid material) there is much less ectoplasm. These changes in the ectoplasm are responsible for the variations in the staining of the bacilli. There are no rigid limits to these staining variations, because there are no fixed variations of the ectoplasm. Acid resistance decreases progressively in saprophytic media, while increasing in parasitic media. The contrary is the case with the capacity to stain blue. The staining of the granules found within the bacilli varies in like fashion; but in order to understand what these granules are, we must describe their mode of formation and their significance.

Granules. All forms of the tubercle bacillus, as they grow older, divide their protoplasm into tiny masses, unequal in volume, which by dehydration progressively draw apart within the bacillus and constitute the resistant forms we term granules. The ectoplasm which coats the surface of the bacilli follows the retraction of the protoplasmic masses
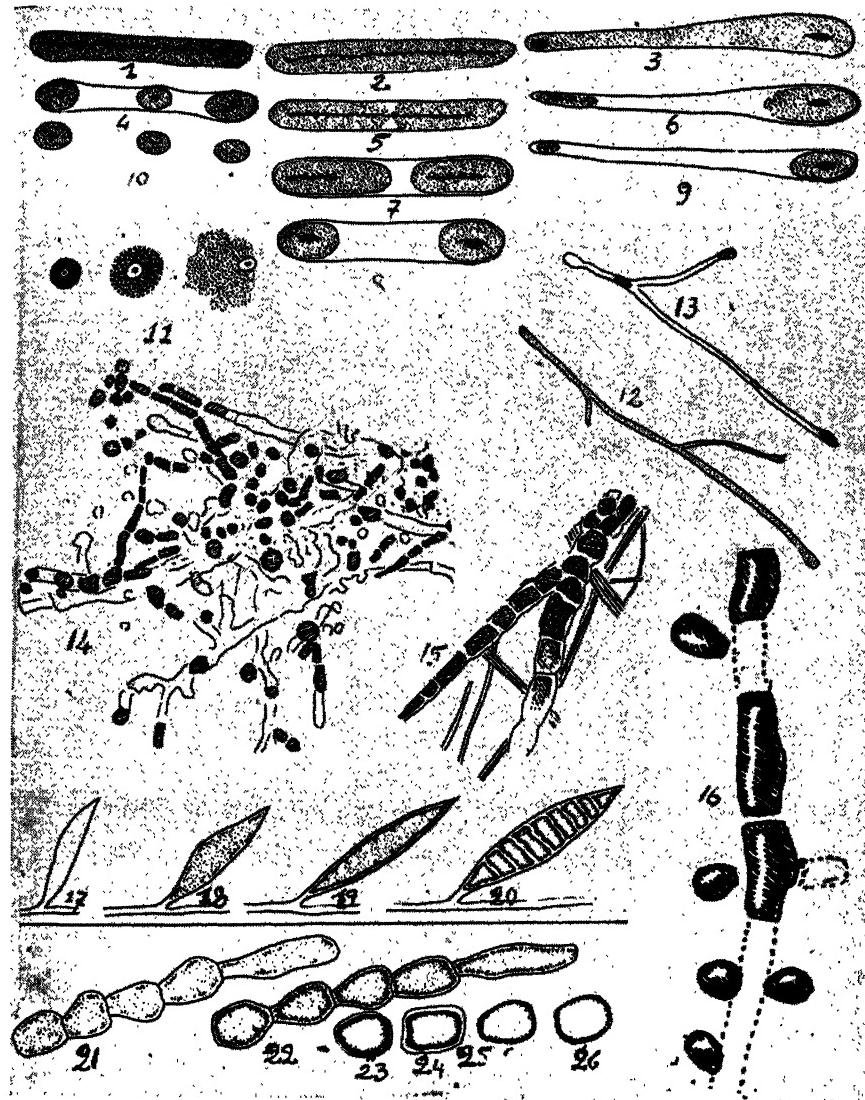
Figure 1. Granule forms and ectoplasm.

1,4,10: Rods of the Koch bacillus (1) with very dense protoplasm. Granules (4), formed by rupture and retraction of the protoplasm. Liberation of granules by rupture of membranes (10).

11: Ectoplasmic lysis, with liberation of plasmatic granules.

2,5,7,8: Larger tubercle bacilli, having less dense protoplasm and two unequal granules.

3,6,9,12,13: Giant forms of the Koch bacillus, with less dense protoplasm, forming small granules separated by spaces because of the greater retraction of the granules.

14,15,16: Dermatophytes of the genus *Achorion* and *Trichopyton* undergoing granular transformation, and showing inequality in size of granules and degree of retraction.

17 to 20: Ectoplasm formation in a *Microsporon*; its creation in the protoplasm (18), and its localization at membrane levels (19-20).

21 to 26: Evolution of ectoplasm in parasitic fungi of the genus *Achorion*.

(Numbers 1 to 13 are diagrammatic.)

within the bacilli, so that the "acid fastness" due to this ectoplasm tends to leave the surface of the bacilli and move towards the surface of the granules within the bacilli. A much mooted question has been the cause of acid fastness—whether it is due to the bacillary wall, or to the proto-plasm of the bacilli. Actually, ectoplasm is the cause; for by moving from the surface of the bacilli towards the surface of the granules which form within the bacilli, it brings about a shift in the acid fastness.

The number of granules varies with each form, and depends on the degree of hydration of the protoplasm. The initial forms which have a denser protoplasm form a larger number of granules, whereas those in the saprophytic stage where the protoplasm is less dense have fewer granules. For the same reason there is variation in the distances between granules. With saprophytic forms the protoplasm is more highly hydrated, the retraction of the granules is more extensive, and the dis-tances separating them are larger. The opposite is the case with parasitic forms, where the protoplasm is denser and the granules draw apart but slightly.

The granules which develop towards the ends of the vegetative forms are larger, because the growth of these forms is always apicular and the protoplasm is more abundant towards the ends of the organisms. Very large end-granules give rise to what we call clubs, which are found in all parasitic fungi under such names as "honeycomb candlesticks," "cellu-lar nails," actinophytes, etc. These clubs have no special significance, apart from the fact that their granules are large. The author's researches on the granules of parasitic fungi have developed the following principles:

Granule volume is controlled by a nucleoplasmic (i.e., nucleus/proto-plasm ratio) constant, and by a cyto-ectoplasmic (i.e., cytoplasm/ecto-plasm ratio) constant, and this leads to a constant ratio between the nucleus, the protoplasmic mass, and the ectoplasm. Virulence depends upon this last named ratio, because it is only by diffusion outside of the bacillus that the endotoxin can become an exotoxin.

According to these principles, the smaller a granule the less virulent it is. Small granules have feeble vitality, so that filtrates containing only smaller granules are not very virulent. It is necessary, therefore, that a filtrate should contain many granules in order that after aggregation and inoculation it can produce the positive reactions which we will describe below.* Consequently, when the bacilli are young and have

---

* After the author had described the process of formation of these granules of the Koch bacillus (*Bull. Acad. de Medicine,* 1934), his lamented friend M. Vaudremer, author of authoritative publications on the Koch bacillus, wrote him a letter from which the following is extracted: "Your work is exciting; nothing is more just than your criticism of the methods heretofore used to study bacilli in general, and tuber-

not undergone granular transformation, inoculation with filtrates is often negative. Furthermore, the methods used in preparing filtrates † adversely affect the granular transformation of the bacilli, thus increasing the number of granules passing through the filter and tending to render inoculation positive.

Multiplication. Among the several modes of reproduction of the tuberculosis bacillus, schizogenesis (duplication by fission) is considered classical; but it seems to be nonexistent because the bacilli do not divide into two like fractions. What gives the appearance of this mode of division is the breaking up of the protoplasm into endocellular granules which are aligned within the bacilli. However, there is no division of the bacilli into two like fractions with the participation of membranes, as we assume in duplication by fission, but a breaking up of the protoplasm into *unequal* masses which, on development within the bacilli, produce granules of unequal volume. This inequality among the granules increases with the saprophytic evolution of forms, but diminishes under the influence of parasitism.

Another mode of division of the Koch bacillus has been described by Miehle, a partial sidewise cleavage, the two bacilli sticking together along part of their length, giving a Y-shaped form. These Y-forms do not come about by sidewise splitting, but arise when two bacilli adhere along part of their length during the initial phase of agglutination, as will be set forth below. Fontes believed that the filtrable granules are the original virus form of the Koch bacillus, and that the bacillus arose from a linear assemblage of several granules. This view appears untenable because no organism shows such a development, and the linearity of the granules is due to the breaking up of the protoplasm into little granules aligned within the bacillus.

Thus, the Koch bacillus does not reproduce by fission, cleavage, or by linear assemblage of granules. There is, nevertheless, a germination of granules such as we observe in the granules or resistant forms (spores) of the other parasitic fungi. How, then, does the bacillus reproduce itself and multiply?

culosis bacilli in particular. To attempt to investigate so delicate a thing as a dead bacillus, stained by incredibly coarse methods, is to block all progress. Following your work, line by line, I see the point you are making, and you must know that I see it with greatest satisfaction.

"I thought, and I think that the term 'granulated' indicates merely great mental satisfaction, and that what is called 'becoming granular' is without a definable inferior limit.

"From the smallest form down to those where pleomorphism is most marked, corpuscles are found always at points corresponding to those where the protoplasmic masses are most condensed; these contract very slowly in the course of sporulation in ordinary cultures.

"Your conclusions I accept naturally, having nothing to say against them . . ."
† See Valtis, J., "Le virus tuberculeux," Paris, Edition Masson et Cie.

Sexual Phase

Agglutination. A most striking fact in the development of the tubercle bacillus is that in ageing cultures, or on loss of vitality, or when the milieu becomes improper for the development of the bacillus, all the various forms are blended together. Kahn and Torrey followed microscopically the development of a single bacillus and observed that it breaks up into granules which, as soon as they become detached, assemble into groups. This agglutinating tendency of all asexual forms explains why in all of our microscopic preparations we see large numbers of bacterial groups which we called "clumps" but to which no significance was attached because according to microbiological concepts bacilli exist only as immutable entities, the agents of microbial disease. There existed among microbiologists an inclination to adhere to fixed ideas, and this denied the notion of transformation. Much can be said about this type of outlook, which has kept men in ignorance over twenty centuries, and has caused dire consequences in all fields of endeavor.

The view that microbes agglutinate in order to undergo lysis and to die appears incredible, for nature does not follow fantastic procedures of this kind; microbes can die quite easily when isolated and without being brought together before dying. It is certain that our theories as to agglutination need justification, for, as Balzac said, too much of our science verges toward negation. Physical chemistry has often served to correct erroneous views. The author was thus led to do research on the problem of the agglutinated masses of microbes, since he considered our theories of agglutination to be inadequate.

"Voiles." Agglutinated bacilli appear at first to be a network of serrated links which, little by little, become compact, turning into what the author called *voiles*. Alexander-Jackson calls them *zoogleal forms* and has published electron-microscopic pictures of them. Besancon and Philibert, Vaudremer, Legroux and Magrou, Haudroy, Arloing, Malatre and Dufourt, Lobstein, Fontes, Karwacki, Kahn and Torrey, have studied in the Koch bacillus what they called *voiles* or *cyanophile reticulum*. In the case of *Bacterium coli*, Winkler called it "plasmodium," and in the case of *Azobacter*, Löhnis called it "symplasm." Descriptions of the mode of formation or the evolution of these forms, as given in texts, are very inaccurate.

The author's researches on *Actinomyces* and the Koch bacillus indicate that the symplasmic matrix is devoid of vitality, because it results from a gelatinization of the ectoplasm which forms the surfaces of the bacilli. This ectoplasmic jelly, probably produced after the agglutination of asexual forms by the action of an autolytic enzyme, covers and protects
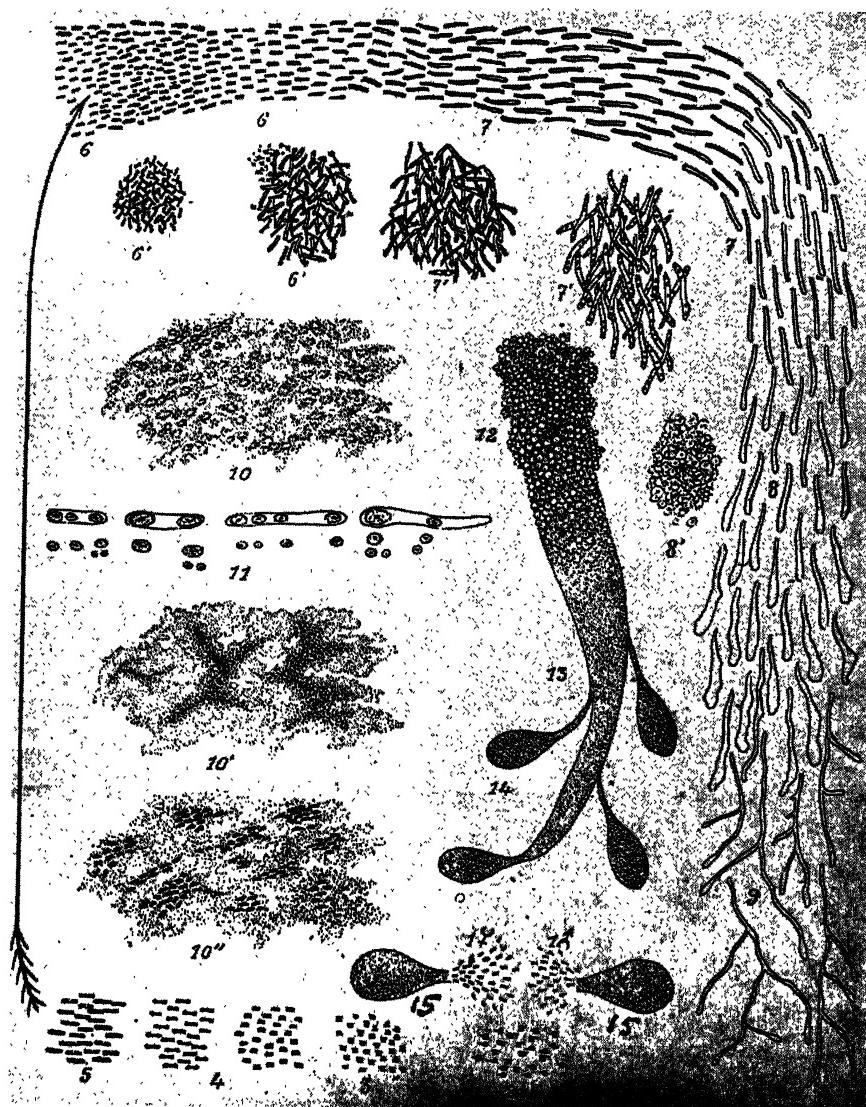
Figure 2. Asexual and sexual cycle of the tubercle bacillus.

6,7,8,9: Successive transformation of asexual diminutive rod forms (6) into non-acid-fast bacilli (7), and giant and ramifying forms (8, 9). Note progressive increase in volume of the bacilli.

6′,7′,8′: The asexual forms agglutinate into masses or clumps to enter the sexual phase.

10,10′,10″: "Voiles" formed by lysis of the ectoplasmic envelopes of the agglutinated bacilli.

11,12: The protoplasmic material separates out in granules (11), which are shown in a state of fusion (12).

13,14: From the granular symplasm there emerges a large filament (13), which produces large ovoid cells (14).

15, 1 to 6: The ovoid cells become detached and eject ciliated and non-ciliated gametes (1 ♀ and 1 ♂) which fuse (2) and produce "diplococci" (3); and these, in turn, by apical growth become diminutive rods (4) and recommence the asexual phase (5 and 6).

the nucleoplasmic material of the bacilli, which has been changed into granules. Furthermore, these *voiles,* at the conclusion of their evolution, shrivel up like dried leaves. The photographs of Alexander-Jackson favor this view. Let us now consider what happens to the granules, which constitute the living matter of the agglutinated bacilli.

Symplasms. Daily examination of living material is necessary in order to distinguish and follow the granules. Above all, examination after the thirtieth day of development is essential, as the evolution of the sexual phase is slow and the granules are difficult to observe because of their small size and their refringence.

To facilitate these investigations, the author used a liquid medium based on 4 per cent glycerin water with one per cent peptone, or else simply potato water, to serve as a saprophytic stock material. Before sterilizing these media, they were filtered to remove all particles coming from dissolved materials or from the potato. These liquid media have the advantage of preserving the various forms from desiccation, and of making the rate of their evolution more uniform. In order to avoid specimens that were too large, needles were used instead of pipettes, the material being spread out in droplets on the slide, covered with the cover glass; the specimen was thus examined while living. In this manner the author followed hundreds of preparations.

All examinations were made on fresh material, that is, on unfixed and unstained preparations. In order to see the granules, high magnification (1000 to 1500 diameters) is necessary, with diaphragm somewhat closed because of the refringence of the material. Alexander-Jackson has used a triple stain on the *voiles* (zoogleal forms); although Grigoraki has not yet worked with this method, her photographs appear to be satisfactory.

In 1938 in the course of the author's observations on living specimens, he was delighted to make out, *after the fortieth day of the development of the cultures,* great numbers of granules which had fused together; and he found that these fused granules formed a flat or spread-out mass with irregular borders, about 8 $\mu \times 15 \mu$. He called this mass of fused granules *symplasm.*

One of the extremities of this symplasm sprouts out a large filament having a homogeneous protoplasm and bordered with a fine stripe, as is seen in embryonic cells. This large filament is about 20 μ long, and 4 to 5 μ wide; and it is probable that investigators who have reported seeing large filaments have actually seen these symplastic forms which have developed large filaments. These symplasms are also to be seen in the electron micrographs of Alexander-Jackson, and Dufourt of Lyon has advised the author that he has occasionally observed analogous forms. Towards the end and the side portions of these large filaments, there
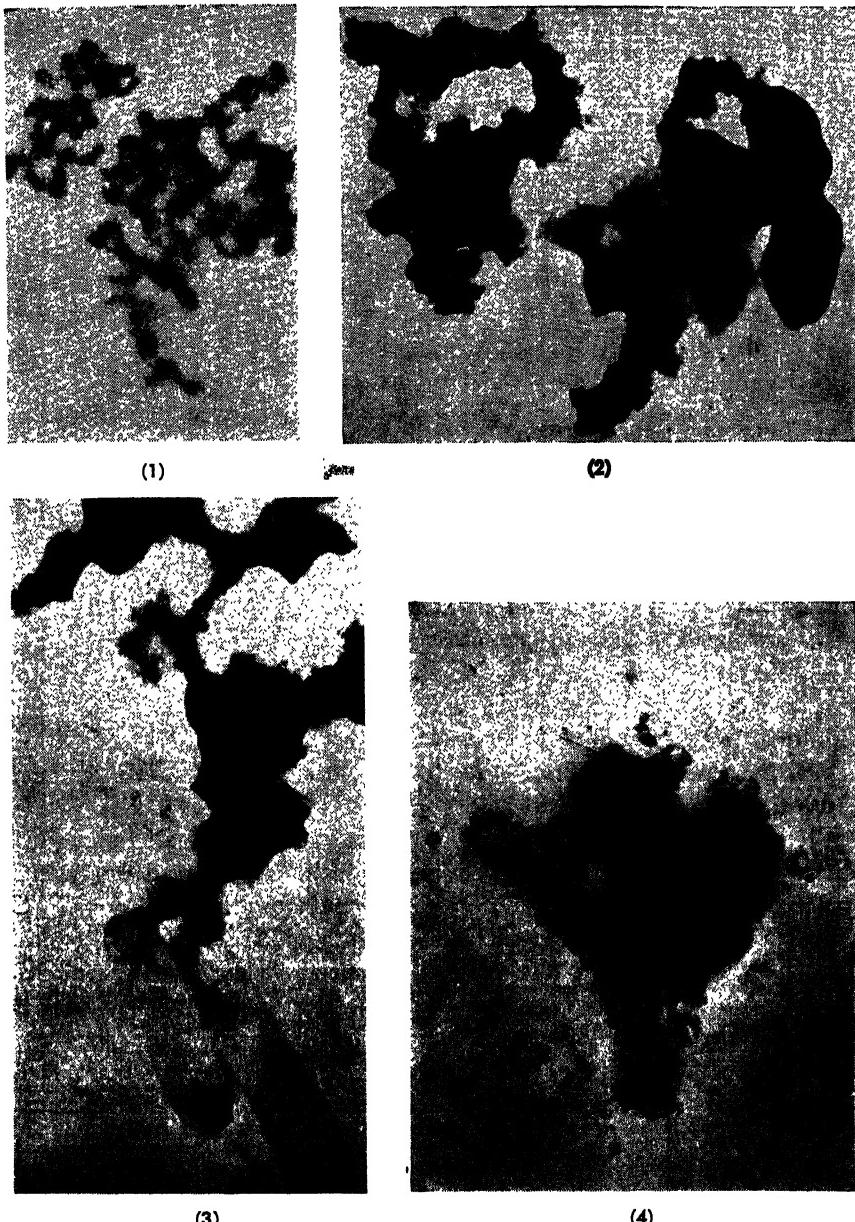
Figure 3. Micrographs of the tubercle bacillus.

(1) Agglutinated bacilli and the beginning of lysis of the ectoplasmic envelopes (light microscope). (2,3,4) Electron micrographs, taken by James Hillier (RCA Laboratories). The products of ectoplasmic lysis form "voiles," which appear dark in the print. The large filaments seen in (3) and (4) probably arise from the symplasm. In (4) may be seen gametes (cocci and diplococci), the "voiles" dissolving, meanwhile, as seen in (3).

develop from a long petiole several *large cells,* ovoid in form, 2 to 3 μ across and 3 to 4 μ long. Reenstierna and other investigators have seen isolated cells like this, and described them as yeast-like forms. These large cells have a fine stripe on their border and their protoplasm, at first homogeneous, becomes granular on ageing. Breaking of the petioles sets these cells free, and at the points of rupture there is seen a conical sheaf composed of *cocci-like forms.* These cocci are of two kinds: one group is round, the other is elongated like a comma. The latter probably correspond to a ciliated form observed by Courmont in specimens made from the products of old tuberculous cavities. Each of these cocci-like forms arises from a different ovoid cellule. Two different cocci fuse together to give a *diplococcus-like form;* complete fusion of the two cocci gives a coccobacillus, which elongates by apical growth and forms a *rod.* This last corresponds to the initial form of the asexual phase, and the cycle recommences.

Conclusions

The tubercle bacillus develops in two phases:

(1) *An asexual phase,* represented by all the vegetative forms whose evolutionary order is as follows: tiny rod, then the bacillary form, and finally giant forms. These transformations occur because of the degradation which the organisms undergo under the influence of nutrition, which increases their volume but decreases their vitality and virulence. On ageing all these forms develop granules within.

(2) *A sexual phase,* which begins by an agglutination of all the asexual forms. This agglutination develops when the forms undergo ageing or become exhausted, or develop under unfavorable conditions. The ectoplasm which covers the surfaces of the agglutinated bacilli gelatinizes and gives an amorphous *voile,* while at the same time the nucleoplasmic substance is transformed into granules within the bacilli. The fusion of these granules when liberated by the lysis of the ectoplasm gives rise to a *symplasm.** This symplasm elongates and develops a large filament carrying large ovoid cellules which break free and discharge *cocciform gametes.* These cocciform gametes fuse two by two and produce tiny rods which recommence the asexual phase.

Phagocytosis and Relapse in Cyclic Diseases. Massive reproduction of gametes at the end of the sexual phase is responsible for the periodic outbursts of the disease, and perhaps the same is the case with relapses in cases of cyclic diseases. There is no explanation for a relapse if we

---

* This behavior recalls what happens in slime molds, whose free-living units aggregate to form a pseudoplasmodium, from which sprouts a sporocarp.—*Ed.*

assume that immunity has brought about the lysis and the death of the
bacilli.

Phagocytes whose phagocytotic index (Achard) increases with agglu-
tination, have the task of ridding the organism of agglutinated masses of
microbes, and not of fighting against them. If these inactive masses are
completely eliminated by the phagocytes before the end of the sexual
phase, there is no relapse. On the other hand, if the agglutinated masses
complete their sexual phase before being eliminated, and thus produce
gametes, these gametes have a higher degree of vitality than the im-
munity created by the first onset of the disease, and a relapse occurs in
an attenuated form.

Heredity and Contagion. Granules of the Koch bacillus can pass
through the placenta and produce a fatal granuloma in the fetus. In the
establishment of the tuberculous syndrome, apart from the factors of
the microbe and the terrain, we must consider the accumulation of
microbes within the organism. There must be a certain minimum
quantity of forms before the bacilli can reach their sexual phase. There-
fore contagion can develop only by prolonged contact.

Classification. In agreement with Grigoraki's teacher, Cuicart, *Myco-
bacterium tuberculosis* must be placed with the *Actinomycetaceae,* for
Actinomycetes also have a sexual phase, as described by the author in
1932.

REFERENCES

Polymorphism:

1. Alexander, E. G., "Developmental Morphology of Human and Avian Tubercle
   Bacilli on Bordet-Gengou Medium," *Proc. Soc. Exptl. Biol. Med.,* **31,** 1103 (1934).
2. Alexander-Jackson, E., "Studies on the Dissociation of Tubercle Bacilli with Special
   Reference to the Avian and Human Types," *Am. Rev. Tuberc.,* **33** (6) 767 (1936).
3. Arloing, F., Dufourt, and Malatre, "Études sur les variations morphologiques et
   pathogènes du Bacille de la tuberculose," *Paris médical,* **22** (Jan. 2, 1926).
4. ———, and Malatre, "Recherches histobacteriennes comparatives sur les cultures
   de Bacille tuberculeux, d'Actinomycesi et d'Aspergillus sur pomme de terre,"
   Soc. de Biol., 27 avr. 1925, *Compt. rend. soc. biol.,* **92,** 1303 (1925).
5. ———, and Dufourt, "Obtention de formes myceliennes du Bacille tuberculeux,"
   soc méd. des Hopitaux de Lyon, 19 mai 1925, *Lyon médical,* **136,** 443 (1925).
6. ———, and ———, "Polymorphisme du Bacille tuberculeux, formes mycéliennes
   ramifiées," *Bull. Soc. d'Et. Scient. sur la Tuberculose,* **6** (3), 517.
7. ———, and ———, "Variations morphologiques du virus tuberculeux données qui
   s'en dégagent," Soc. med. des Hopitaux de Lyon, 31 mai 1927, *Lyon médical,*
   140, 463 (1927).
8. Arloing, S., "Sur l'obtention de cultures et d'émulsions homogènes du Bacille de la
   tuberculose," Acad. des Sciences, 9 mai 1898, *Compt. rend.,* **126,** 1319 (1898).
9. ———, "Variations morphologiques du Bacille de la tuberculose de l'homme et des
   mammifères obtenues artificiellement," Acad. des Sciences, 20 janv. 1908, *Compt.
   rend.,* **146,** 100 (1908).
10. ———, "Variabilité du Bacille de la tuberculose," *Revue de la tuberculose,* **5** (2),
    1 (1908).

11. ———, and Chantre, "Sur les variations morphologiques et pathogéniques de l'agent de l'infection purulente chirurgicale," *Arch de Physiol.*, 7 (5), 610 (1895).

12. Babes and Levaditi, "Sur la forme actinomycosique du Bacille de la tuberculose," *Arch. de Médecine expérim.*, 9 (1), 1041 (1897).

13. Bauverie, "Études sur le polymorphisme des Champignons, influence du milieu," Thèse de Sciences nat., Lyon, 1900.

14. ———, "Polymorphisme et place dans la classification de *l'Azotobacter chroococcum*," Assoc. franc. pour l'avancement des Sciences, 49e session, Grenoble, 1925, *Compt. rend.*, 361 (1925).

15. Boudin, "Les variations morphologiques du microbe de la tuberculose," Thèse de Med., Paris, 1906.

16. Calmette, A., "Sur la question du transformisme des Bacilles paratuberculeux et des Bacilles tuberculeux," Rapport à la 4e Conference de l'Union Internat. contre le tuberculose, Lausanne, 1924, *Bull. Inst Pasteur*, 22, 593 (1924) and *Revue de la tuberculose*, 5 (3), 716 (1824).

17. Coppen Jones, A., "Ueber die Morphologie und systematische Stellung des Tuberkelpilzes und uber die Kolbenbildung bei Aktinomykose und Tuberkulose," *Centr. für Bakt.*, 17 (I), 1-70 (1895).

18. Courmont, J., and Nicholas, J., "Sur une tuberculose steptobacillaire d'origine bovine," *Arch. de Parasitol.*, 1, 123 (1898).

19. Ferran, J., "Note relative aux aptitudes saprophytiques du Bacille de la tuberculeux," Acad. des Sciences, 2 oct. 1897, *Compt. rend.*, 125, 515 (1897).

20. Grigoraki, L., "Morphologie et cytologie des Actinomyces," Acad. des Sciences, 5 oct. 1931, *Compt. rend.*, 193, 540 (1931).

21. ———, "De l'action de l'ether sur les micro-vegetaux parasites du tissu animal (facteurs de virulence, vitalite, degradation et mutation), "Acad. des Sciences, 25 sept., *Compt. rend.*, 195, 555.

22. ———, "Morphologie et cycle évolutif des Actinomyces," Acad. des Sciences, 15 fev. 1932, *Compt. rend.*, 194, 641 (1932).

23. ———, "Senescence et evolution," Vol. 1, Chap. 3, p. 170, Paris, Edit. Le Francois, 1944.

24. Kahn, M. C., "A Developmental Cycle of the Tubercle Bacillus as Revealed by Single Cell Studies," *Am. Rev. Tuberc.*, 20, 150 (August, 1929); *Bull. Inst. Pasteur*, 27, 855 (1929).

25. ———, and Torrey, J. C., "The Developmental Cycle of the Tubercle Bacillus as revealed by Single, Cell Studies," *Am. Rev. Tuberc.*, 18, 815 (Dec., 1928); *Bull. Inst. Pasteur*, 27, 289 (1929).

26. Karwacki, L., "Mutations des Bacilles tuberculeux en Streptothrix," Soc. polonaise de Biol., 20 juin 1928, *Compt. rend. soc. biol.*, 99, 1152 (1928).

27. ———, "Saprophitisme des Bacilles tuberculeux," Soc polonaise de Biol., 27 févr. 1929, *Compt. rend. soc. biol.*, 100, 1152 (1929).

28. ———, "A propos de certains caractères de culture de Streptothrix obtenus par transformation de bacilles acido-résistants," Soc. pol. de Biol., 27 juin 1928, *Compt. rend. soc. biol.*, 99, 1173 (1928).

29. Kirchensteins, "Sur la structure et le mode de développement du Bacille tuberculeux," *Ann. Inst. Pasteur*, 36, 416 (1922).

30. ———, "Sur la morphologie et le mode de développement des formes atypiques des Bactéries," Soc. de Biol., de Lettonie, 23 sept. 1922, *Compt. rend. soc. biol.*, 88, 716 (1923).

31. Petit, "Contribution a l'étude cytologique et taxinomique des Bactéries," Acad. des Sciences, 15 mars 1926, *Compt. rend.*, 182, 717 (1926).

32. Reenstierna, "Sur la vie non acido-résistante du parasite de la tuberculose," *Arch. Inst. Pasteur de Tunis*, 15, 309 (1926).

33. Vaudremer, A., "Le Bacille de la tuberculose Sclerothrix Kochii," *Paris médical*, 26 (Jan. 5, 1924).

34. ———, "Recherches sur le developpement du Bacille tuberculeux et la vaccination

antituberculeuse du cobaye," Acad. de Medecine, 4 mai 1926, *Bull.*, 95 (3), 440 (1926).

35. ———, "Un Bacille tuberculeux humain, un Bacille tuberculeux bovin acido-résistants facultatifs," Soc. de Biol. 5 fév. 1921, *Compt. rend. soc. biol.*, 84, 259 (1921).

Granule Forms and Filtrable Virus:

1. Arloing, F., and Dufourt, "Résultats obtenus par l'inoculation de virus tuberculeux filtrant aux cobayes nouveau-nés," Soc. de Biol., 4 déc. 1926, *Compt. rend. soc. biol.*, 95, 1414 (1926).
2. ———, and ———, "Le virus tuberculeux filtant et les acquisitions nouvelles sur l'hérédité tuberculeuse," *Presse médical*, 17 (Jan. 5, 1927).
3. ———, and ———, "Le virus tuberculeux filtrant et son passage à travers le placenta," Acad. de Médecine, 16 févr. 1927, *Bull.*, 95 (3), 163 (1927).
4. ———, ———, and Malatre, "Les formes filtrantes du Bacille de Koch et le problème de l'hérédité tuberculeuse," Soc. méd. des Hopitaux de Lyon, 19 mai 1925, *Lyon médical*, 136, 144 (1925).
5. Calmette, A., "Les éléments filtrables du virus tuberculeux," Rapport à la 6e Conference de l'Union Internat. contre la tuberculose, Rome, 22-26 sept. 1928, *Bull. Inst. Pasteur*, 26, 889 (1928).
6. ———, and Valtis, J., "Les éléments virulents filtrables du Bacille tuberculeux," *Ann. de Médecine*, 19, 553 (1926).
7. Fontes, "Saprophytismo do Bacillo tuberculoso," Rapport présenté au 2e Congrès Pan-Américain de la tuberculose, Rio de Janeiro, 1929, *Mém. inst. Oswaldo Cruz*, suppl. 205 (Aug. 31, 1929).
8. Grigoraki, L., "La sporulation du Bacille tuberculeux et sa forme filtrable," Académie de Médecine, 6 mars 1934, *Bull.* (1934).
9. Hauduroy, "État actuel de la question des formes filtrantes du Bacille tuberculeux," *Presse médical*, 227 (Feb. 20, 1926).
10. ———, and Vaudremer, "Recherches sur les formes filtrables du Bacille tuberculeux," Soc. de Biol., 22 déc. 1923, *Compt. rend. soc. biol.*, 89, 1276 (1923).
11. Hauduroy, "Les ultravirus et les formes filtrantes des microbes," Paris, Edit. Masson et Cie., 1929.
12. Nicolle, Ch., "Sur la nature des virus invisibles, origine microbienne des infra-microbes," *Arch. Inst. Pasteur de Tunis*, 14, 105 (1925); *Revue d'Hygiène*, 47, 408 (1925).
13. Valtis, J., "Le Virus tuberculeux," Paris, Edit. Masson et Cie., 1932.
14. Vaudremer, A., "Formes filtrantes des Bacilles tuberculeux," Soc. de Biol., 9 juin 1923, *Compt. rend. soc. biol.*, 89, 80 (1923).
15. ———, "Reflexions sur la filtration des Bactéries," Soc. de Path. comp., 8 février 1927, *Revue de Path. compar.*, 482 (1927).

Sexual Forms:

1. Alexander-Jackson, E., "A Hitherto Undemonstrated Zoogleal Form of Mycobacterium Tuberculosis," *Ann. New York Acad. Sci.*, XLVI, Art. 2, 127-152 (1945).
2. Arloing, S., and Courmont, P., "De l'obtention de cultures du Bacille de Koch les plus propices à l'étude du phénomène de l'agglutination," Acad. des Sciences, 8 août 1898, *Compt. rend.*, 127, 312 (1898).
3. Besancon, and Philibert, "Etude sur la structure des voiles du Bacille tuberculeux," Soc. d'Et. Scient. sur la Tuberculose 12 mars 1914, in Bull., 4(a), 32 (1914); *Presse médicale*, 327 (1914).
4. ———, ———, and Hauduroy, "Sur la structure des voiles jeunes des cultures de bacilles tuberculeux," Soc. de Biol. 23 février 1924, *Compt. rend. soc. biol.*, 90, 475 (1924).
5. Grigoraki, L., "Morphologie et Cycle évolutif des Actinomyces," Acad. des Sciences, 15 févr. 1932, *Compt. rend.*, 194, 641 (1932).

6. ——, "De l'action de l'éther sur les micro-végétaux parasites du tissu animal (facteurs de virulence vitalité, dégradation et mutation)," Acad. des Sciences, 25 Sept., 1932, *Compt. rend.*, **195**, 555 (1932).

7. ——, "Senescence et Evolution," Vol. 1, Chap. 3, p. 170, Paris, Edit. Le François, 1944.

8. Hauduroy, "Toxines diphtériques donnant naissance à un Bacille diphtérimorphe," Acad. des Sciences, 14 fév. 1927, Compt. rend., **184**, 409 (1927).

9. Legroux and Magrou, "Etat organisé des colonies bactériennes, *Ann. Inst. Pasteur*, **34**, 417 (1920).

10. Löhnis, F., "Zur Morphologie und Biologie der Bakterien," *Centr. für Bakt.*, **56** (II), 529 (1922).

11. ——, and Smith, N. R., "Studies upon the Cycles of the Bacteria," Part II, Life cycle of Azotobacter, *J. Agric. Research*, **23**, 401 (1923); *Centr. für Bakt.*, also cited by Bauverie (270) **62** (II) (1924).

Biochemical, Clinical, Classification:

1. Achard, Les Maladies Cyclique," Vol. I, Paris, Masson et Cie., 1932.

2. Besancon, and Philibert, "Sur la nature de l'acido-résistance du bacille tuberculeux," *Bull. Soc. d'Et. Scient. sur la tuberculose*, 214 (Nov. 14, 1907); *Bull. médical*, **21**, 1081 (1907).

3. ——, and ——, "Le Bacille acido-résistant n'est qu'une des formes du parasite de la tuberculose," *Presse médicale*, **33** (Jan. 9, 1926).

4. Lobstein, "Recherches biochimiques sur le Bacille tuberculeux," Thèse de Sciences nat., Strasbourg, 1922.

5. Löhnis, F. and Freed, "Textbook of Agricultural Bacteriology," New York, McGraw Hill Book Co., Inc., 1923.

6. Roman, E., "Le Bacille tuberculeux," Thèse de Lyon, 1930.

7. Vaudremer, A., "Le Bacille tuberculeux, études bactériologiques, cliniques et thérapeutiques," Paris, Presses Universitaires de France, 1927.

8. ——, "Tuberculine et miliéux de culture du Bacille tuberculeux," Soc. de Biol., 30 avr. 1921, *Comp. rend. soc. biol.*, **84**, 775 (1921).

9. Guiart, J., "La place du bacille tuberculeux et des autres Bactéries dans la classification des Champignons," *Lyon médical*, **140**, 469 (Nov. 13, 1927).

10. Howard, "Phagocytose, lyse et perte de l'acido-résistance du Bacille de Koch en présence des leucocytes de cheval immunisé," Réunion biologique de Buenos-Aires, 8 sept. 1922, *Compt. rend. soc. biol.*, **87**, 1054 (1922).

11. Dufourt, "Traité de Phtisiologie chimique," Paris, Edit. Vigot frères, 1945-46.

12. Grigoraki, L., "Les oeufs d'oursins et la tuberculose," *Marseille-médical*, 261-263 (1935).

13. ——, "L'oxygène et l'évolution du bacille tuberculeux," *Lyon médical*, No. 21 (May 26, 1946).

EDITOR'S NOTE

The following criteria are mentioned by Dr. T. M. Rivers as being in use to distinguish *viruses* from *bacteria:*

(1) Viruses pass through bacteria-tight filters. (However, filtration is not always reliable)

(2) Viruses are, for the most part, not visible in the ordinary microscope. [However, there is a gradual transition in size from viruses approximating molecular dimensions (e.g., foot-and-mouth disease virus, 10 mμ) up to sizes approximating those of small bacteria (e.g., psittacosis virus, 275 mμ; Rickettsia, 300 mμ; *Bacillus prodigiosus*, 750 mμ). See W. M. Stanley, Vol. V of this series (1944), p. 807; and "Life, Its Nature and Origin," by J. Alexander, p. 81, (Reinhold Publishing Corp., 1948).]

(3) Viruses can grow only in the cells of a susceptible host. [A recent report states that the Rous "virus" has been cultivated in cell-free artificial media—see V. Wuerthele-Caspe, *et al., J. Am. Med. Women's Assoc.*, **4**, 135-41 (1949)].

Furthermore, evidence is accumulating that some bacteria may give rise to forms of viral dimensions, which may regenerate the larger forms. See, e.g., "The Filtrable Forms of Bacteria: I. A. Filtrable Stage in the Life History of the Shiga Bacillus," by Philip Hadley, Edna Delves, and John Klinek, *J. Infectious Diseases*, 48, 1-159 (1931).

The principles of general semantics (A. Korzybski) should be invoked here. Obviously, the particular classification or name applied to a microbial entity will not affect what it actually does to the cells and tissues of a susceptible host. Facts should dominate the progress of science, not sterile wrangles over nomenclature.

More recently, a mycobacterium-like pleomorphic organism has been isolated from many kinds of cancers, and has been grown in serial cultures *in vitro*. Hofrat Prof. Dr. Franz Gerlach (Möding-Wien) in his book "Krebs und obligater Pilzparasitismus" (Urban and Schwarzenberg, Wien, 1948), describes his long-continued and extensive work with this organism, referring on p 28 *et seq.* to the observations of Prof. Nello Mori (Naples), whose paper "Hypothèse sur la nature microbienne du cancer" appeared in *L'Hygiéne Sociale*, 101, 2032 (1933), with later papers in *Reforma med.* (1937-38). Mice inoculated by Gerlach with the organism cultured from cancers all developed some form of abnormal or diseased condition (cachexia, viscous exudates, etc.), though only 4.12 per cent developed actual tumors. Gerlach expresses the view (p. 127) that cancer is a generalized infection, the tumor being a local manifestation of it, and so to say its final and most evil outcome. This view accords with the fact that on inoculation into animals the organism becomes widely distributed and may be recovered again in cultures.

Entirely unaware of Gerlach's work, Dr. Virginia Wuerthele-Caspe (Newark, N. J.), in studying cases of scleroderma, observed an acid-fast organism in patients and sought the cooperation of Dr. Alexander-Jackson, who had done extensive work on the pleomorphic forms of *M. tuberculosis* [see e.g., "A Hitherto Undemonstrated Zoogleal Form of Mycobacterium Tuberculosis," *Ann. N. Y. Acad. Sci.*, 46, 127-152 (1945)]. In her paper Dr. Alexander-Jackson pointed out the similarity of pleomorphic forms of *M. tuberculosis* to pleuropneumonia-like and other forms described in other organisms [see "The Significance of Pleomorphism in Bacteroides Strains," L. Dienes and W. E. Smith, *J. Bact.*, 48 (2), 125 (1944)]. Using a special culture medium, Alexander-Jackson succeeded in cultivating the organism found in scleroderma, although it did not grow well on ordinary media; and these cultures produced characteristic lesions in animals. Her preliminary paper on the cultivation of a mycobacterium from the blood of patients suffering from leprosy (Hansen's disease) has just (June, 1950) been sent to *Science* for publication. Pure cultures of these organisms have produced lesions in mice (alopecia, skin and ear ulcerations).

On examining smears from cancer patients, stained by Alexander-Jackson's triple stain [*Science*, 99, 307-8 (1944)], Wuerthele-Caspe observed small acid-fast forms resembling mycobacteria [see V. Wuerthele-Caspe, *et al.*, *Bull. N. Y. Microscopical Soc.* (August, 1948)]. In cooperation, these two investigators cultivated mycobacterium-like pleomorphic organisms from many cancers. Preliminary skin tests and serological tests carried out with the cooperation of Drs. Ralph di Falco and John Anderson (Rutgers University) suggest the specificity of this organism, in agreement with Gerlach.

Drs. A. E. Gessler, Kenneth S. McCarty, M. C. Parkinson, *et al.*, in a series of papers [see, e.g., *Exptl. Med. and Surgery* (Nov., 1947, Nov., 1948, Nov., 1949)] observed in the electron microscope, in sections made from human cancer tissue, submicroscopic particles not to be seen in similar sections made from normal tissues. The particles were isolated by a special technique; their size ranged from about 20 to 200 mμ, but their shape was always spherical. These particles were indistinguishable from those isolated by the same technique from tumors in ducks induced by the 14D-7 duck variant of the Rous chicken sarcoma virus; and this similarity in size and shape suggested that the particles seen in human cancer may be a causative agent. These same investigators further found that when *d*-tryptophan was substituted in casein hydrolysate for the *l*-tryptophan normally present in the mouse diet, the growth of spontaneous C3H mouse mammary tumors was greatly retarded. They also found that

the same tumors were eradicated by high frequency radiation of 200 to 3,000 mega-
cycles [*Exptl. Med. and Surgery* (July, 1950)].

Dr. Peyton Rous has stated [*Am. Scientist*, **34**, 329-58 (1946)]: "There is evidence
showing that viruses persist in cells as harmless symbionts, ready to bring about disease
when favorable conditions permit. An example of these so-called 'latent' viruses occurs
in the King Edward potato, where it causes no evident disease; but it can kill other
varieties upon inoculation. And a virus infection superimposed upon a 'precancerous'
irritation may cause rapid cancer development." The behavior of J. J. Bittner's "milk
factor" comes to mind.

The newly discovered polymorphic cancer organism has, no doubt, been visible in
some of its various forms to many investigators, but has been generally ignored as an
"artifact" or a "tissue element," since it found no place in the classical textbook
picture. The fact that it can be cultivated *in vitro*, apart from living cells, means
that it is not a virus within the scope of present classifications. The view is slowly but
surely developing that many so called "viruses" and "ultrafiltrables" are either tiny
organisms, or else are submicroscopic forms of pleomorphic organisms, which can and
often do assume larger, microscopically visible forms. *Chemical and Engineering News*
(June 5th, 1950) in reporting a recent address of Prof. Wendell M. Stanley (California)
said: "The larger viruses have a chemical composition and properties which are char-
acteristic not of molecules, but of organisms." They thus form "a new link between
the molecules of the chemist and the organisms of the biologist."

J. A.

BLOOD GROUPS, WITH SPECIAL REFERENCE TO THE Rh-Hr FACTORS

Alexander S. Wiener

Serological Lab., Office of Chief Medical Examiner, New York, N. Y.

IN THE FIFTH VOLUME of this series, Boyd * has presented an excellent survey of general immunology from the chemical standpoint. The present section on blood groups deals with a small but important portion of the field of immunology. The field of blood grouping was opened by the investigations of Landsteiner, who also fathered the field of immuno-chemistry. For his fundamental research, particularly the discovery of the blood groups, Landsteiner was awarded the Nobel prize in medicine in 1930. From the outset he had visualized an individuality of the blood somewhat analogous to the individuality of fingerprints, and this concept has received considerable support recently, especially through the discovery of the Rh-Hr factors in which Landsteiner again played one of the leading roles.[1, 2]

Aside from their practical importance, the blood groups are of interest to the immunologist because they provide one of the simplest and "prettiest" examples of serological reactions. The red cells form stable suspensions which give sensitive and consistent reactions and are therefore ideal for the investigation of immunological phenomena. The blood groups are of interest not only to the serologist but also to the geneticist, because they provide the most striking example of Mendelian inheritance in man, and this knowledge has found practical application in medicolegal cases of disputed parentage. Their most important applications are in clinical medicine, where knowledge of the blood groups and rhesus factors has made blood transfusion safe and has led to the explanation of a hitherto obscure blood disease of fetuses and newborn infants.

General Principles

To determine an individual's blood group, a suspension of his blood cells is tested with a battery of antisera, each of which is specific for a

* Boyd, William C., in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. V, pp. 957-9, New York, Reinhold Publishing Corp., 1944.

different agglutinogen.[3] For example, with the aid of the two sera, anti-A and anti-B, human blood can be tested to determine whether it contains either, both, or neither of two corresponding agglutinogens, A and B, and in that way human blood can be classified into four groups A, B, AB, and O. This scheme of four groups is increased to six groups with the aid of a special serum which distinguishes two main varieties of agglutinogen A. In a similar manner, with the aid of two antisera, anti-M and anti-N, a person can be classified as belonging to one of the three types, M, N, or MN. Recently,[4, 5] another antiserum anti-S, related to M-N system has been found which increases the number of M-N types to six. A third system of blood types has been distinguished with the aid of the three Rh antisera, anti-Rh_0, anti-rh', and anti-rh''; and two Hr antisera, anti-hr' and anti-hr'', which theoretically in combination differentiate as many as eighteen types of blood, some of which, however, are quite rare.[6] Since everybody must belong to one of the A-B-O groups, to one of the M-N types, and to one of the Rh-Hr types, in all, $6 \times 6 \times 18$ or 648 different types of human blood can be distinguished with these reagents. When we take into account other less readily available reagents such as anti-P, anti-rh^w, anti-Le, etc., it is apparent that the number of possible subdivisions is increased to tens of thousands, so that Landsteiner's concept of an individuality of the blood is indeed close to realization.

With the possible exception of the A-B-O substances, the chemical nature of the structures on the surface of the red cell responsible for the blood group reactions is unknown; thus far, these substances have been distinguished by their serological reactions alone. The antibodies in the serum, like other antibodies, are globulins, and the specificity of their reactions is usually explained on a steric basis, the shape of the antibody molecule being conceived as the counterpart of the shape of the corresponding agglutinogen.[7-13] While there seems no doubt that electrostatic forces play a role in these, as well as in other antigen-antibody reactions, it is doubtful, as Landsteiner,[11] Boyd, and others have pointed out, that true covalent bonds are formed.

The polar groups probably initiate the combination of antigen and antibody, while the firmness of the union is most likely maintained by van der Waals forces. The hapten groups on the red cells and the corresponding active patches in the antibody molecules are relatively small, and it may seem strange that a tiny antibody link should be strong enough to hold together relatively massive objects like red cells. The probable explanation is that the hapten groups are repeated at regular intervals about the periphery of the red cells, so that multiple bonds are formed at each area of contact between cells.[14] For example, in Figure 1 we have presented a theoretically possible arrangement of the A-B-O,

M-N, and Rh-Hr haptens on the surface of the erythrocytes. The serological reactions suggest that the A-B-O haptens are far more numerous than the M-N and Rh-Hr haptens. For example, although the reactions in Rh tests correctly performed are just as distinct as in A-B-O tests, the clumps are much more fragile in Rh tests and are easily broken apart by shaking.

In general, agglutination occurs when the specific antibodies link together the corresponding haptens on adjacent red cells. According to this concept the antibodies producing agglutination must be at least bivalent and may be multivalent. Since it is postulated that multiple links are formed, the question may be raised as to how it is possible for the

Figure 1. Diagrammatic visualization of a possible arrangement of haptens on the surface of red cells.

● - A-B-O Haptens
x - M-N Haptens
⊗ - Rh-Hr Haptens
• - Haptens of other specificities;
      species specific haptens
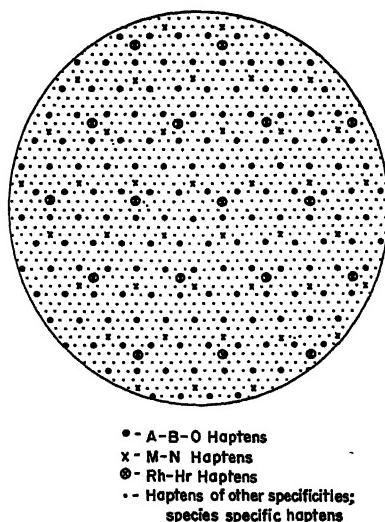
haptens on two contiguous cells to be aligned so that they match precisely. A possible explanation is that the haptens are arranged in a regular pattern on the surface of the cells, so that when any two pairs of haptens on contiguous cells come in contact, all the haptens become aligned. The principle would then be similar to the one utilized when mounting a tire on the wheel of an automobile. It seems apparent therefore that there must be many agglutinogens on the surface of the red cells which cannot be detected serologically. For example, an agglutinogen which is distributed at wide intervals about the periphery of the cells could not be detected, because even if an antibody were formed, the number of links would not be sufficient to hold the red cells together firmly enough to give a distinct reaction.

The recent work on the Rh-Hr system of blood factors had led to the discovery that sensitized (or immunized) individuals form more than one

variety of antibody of a given specificity. When it was found that many individuals with clinical evidence of Rh sensitization had no demonstrable Rh agglutinins in their sera, it occurred to the writer that a different type of Rh antibody might exist, incapable of clumping cells in tests as ordinarily carried out in saline media. Such antibodies were visualized as being univalent instead of bivalent, so that when they combine with red cells in saline media, they "coat" the erythrocytes without
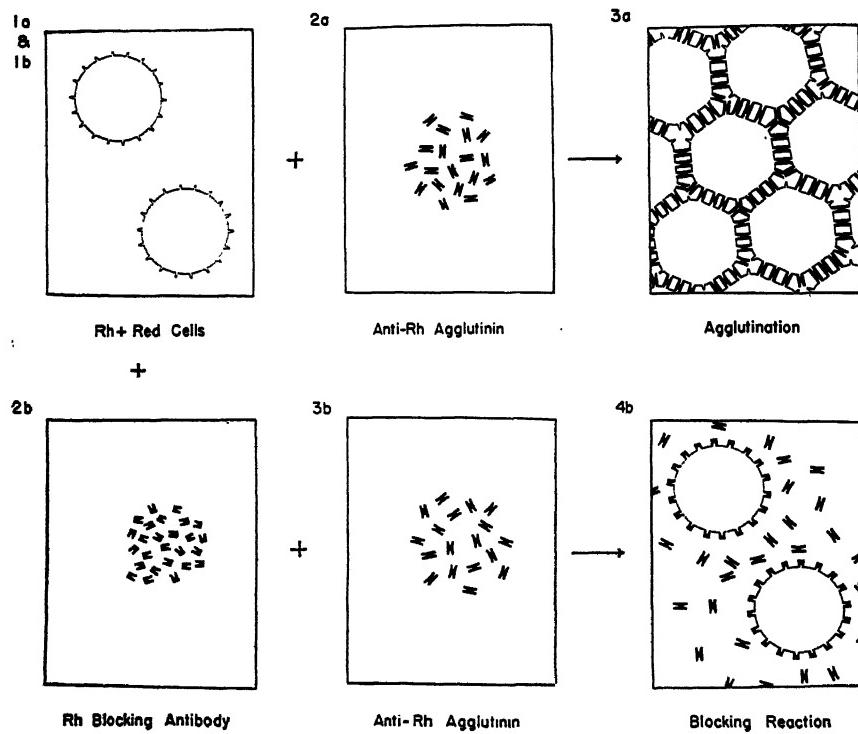
Figure 2. Diagrammatic representation of the reactions of univalent and bivalent Rh antibodies.

clumping them. Based on this concept, the blocking test for detecting univalent antibodies was devised.[15] * It was found, moreover, that when the tests with such antisera were carried out in plasma media or serum media instead of saline, clumping resulted.[16] This is explained by postulating that serum contains a substance called conglutinin, a colloidal aggregate of serum proteins which, like complement, is adsorbed only by specifically sensitized red cells and causes them to stick together. This type of clump-

* Working independently, Race [15a] encountered the Rh-blocking antibody when he pooled Rh antisera in an attempt to produce a polyvalent serum. When the pooled serum failed to give the expected reactions, he traced this to the effect of Rh-blocking antibody in the serum.

ing is called conglutination to distinguish it from the clumping produced by bivalent antibodies, which is designated as agglutination. The differences between the agglutination, conglutination, and the blocking reactions are represented in Figures 2 and 3.
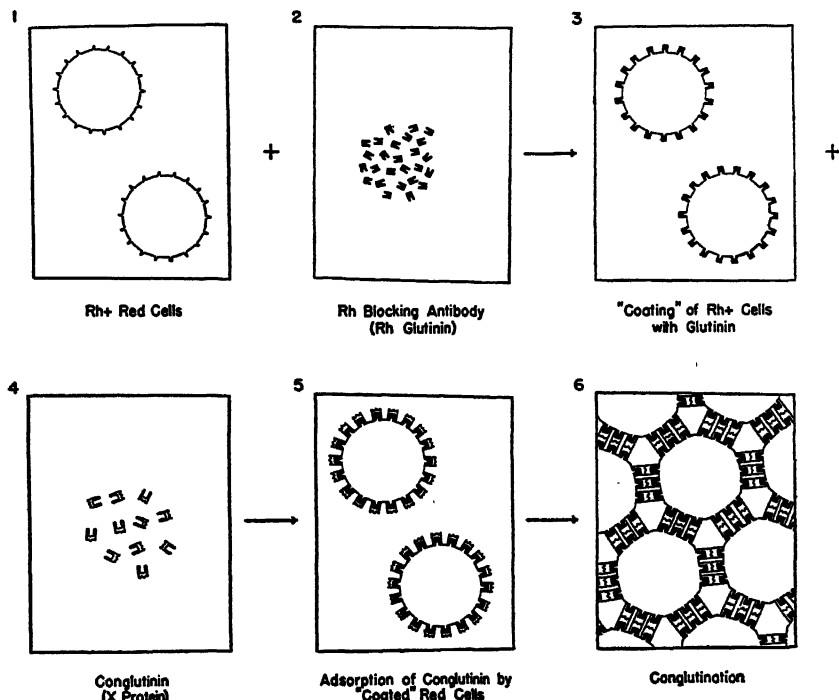
Figure 3. Diagrammatic representation of the conglutination reaction.

Blood Groups and Transfusions

The most important practical application of blood grouping is in the selection of donors for blood transfusion purposes. For this work, the most important blood groups are the original four Landsteiner groups, O, A, B, and AB. These groups are unique in that the isoagglutinins anti-A (or alpha) and anti-B (or beta) exist preformed in normal human sera. On the other hand, normal human sera hardly ever contain preformed agglutinins for the M-N and the Rh-Hr systems of antigens; these reagents can, however, be obtained by immunization.

Under Landsteiner's law, those isoagglutinins are present in the serum and only those for which the corresponding agglutinogen is lacking from the red cells, except in newborn infants. For example, the serum of individuals of group A contains beta-agglutinins but never alpha-agglutinins.

Therefore, if an individual of group A is given a transfusion of blood containing the agglutinogen B (group B or AB) the donor's cells may be clumped and/or lysed in the patient's circulation, giving rise to a serious reaction. In patients surviving the reaction, the titer of the corresponding isoantibody usually rises as a result of the isoimmunization, so that if some time in the future the mistake were repeated, the recipient would probably not be as fortunate the second time. Such dangerous reactions were largely prevented when blood grouping came into routine use for the selection of donors for blood transfusion. During the fourth decade of the present century, with the introduction of blood banks and the more widespread use of the citrate method of transfusion, the number of blood transfusions given increased tremendously. It was then noticed that even when the patient and the donor belonged to the same group, hemolytic reactions occasionally occurred. In a number of these cases it was possible to demonstrate in the serum of the recipient an irregular agglutinin which appeared to be responsible for the reaction. However, no attempt was made to correlate these observations with one another, and no further progress was made until the discovery of the Rh factor was announced by Landsteiner and Wiener in 1940.

Actually, the discovery of the rhesus factor dates back to 1937, when Landsteiner and Wiener [17] conducted an investigation on the evolution of the M agglutinogen. They were able to demonstrate a step-like evolution of this property from the lowest primates up to men. When Landsteiner and Wiener found that rhesus monkeys possessed an M-like agglutinogen, they injected rhesus blood into rabbits, and in that way produced antisera capable of clumping specifically human red cells containing the M agglutinogen. Previously, Schiff and Adelsberger [18] had found that the injection of sheep blood into rabbits resulted in the production of antisera specific for human A blood. It then occurred to Landsteiner and Wiener that it might be possible to produce other heterogenetic antisera which could detect hitherto unknown properties of human blood. Continuing this line of investigation, they then found that certain rhesus antisera contained antibodies for human blood differing in specificity from any antiserum previously described. The rhesus antisera clumped the bloods of about 85 per cent of all the Caucasoids, between 90 and 95 per cent of all Negroids, and practically all Mongoloids. Human bloods clumped by the anti-rhesus sera evidently contain an agglutinogen related to one present in rhesus blood, and are therefore said to be rhesus-positive (or Rh-positive), while bloods not clumped by anti-rhesus serum are said to be rhesus-negative (or Rh-negative).

At first, this finding appeared to be only of academic interest. In 1939, however, Wiener and Peters [19] studied three cases of intragroup hemolytic transfusion reactions, one fatal, and demonstrated in the sera of all three

patients agglutinins of parallel specificity, which clumped the bloods of 85 per cent of all the Caucasoids. This suggested that the blood factor in question might be identical with or related to the Rh factor previously discovered by Landsteiner and Wiener. Indeed, it was found that all three patients were rhesus-negative, while all the incompatible bloods were rhesus-positive, and that the irregular agglutinins gave reactions parallel to the anti-rhesus sera.

It has been found that fully 90 per cent of intragroup hemolytic reactions could be explained on the basis of Rh sensitization, and that if Rh-negative patients are transfused with Rh-negative blood such reactions could be prevented.[19, 20] Since serum of Rh-negative persons does not contain preformed Rh antibodies, transfusions of Rh-positive blood are not harmful until such recipients have become sensitized. It has been found that Rh sensitization could arise in one of two ways, namely, as a result of a previous injection of Rh-positive blood or a pregnancy with an Rh-positive fetus.[19, 21]

As a result of these findings, Rh testing has become a routine preliminary procedure before all blood transfusion procedures, and all Rh-negative individuals are given Rh-negative blood, not merely to avoid a reaction, but also to avoid the development of Rh isosensitization. As will be explained in a later section, aside from its role in transfusion, Rh sensitization of an Rh-negative prospective mother may cause a blood disease in her Rh-positive fetus, as was first demonstrated by Levine and his collaborators.[22]

An important practical problem is the production of Rh testing serum. While much of the original work was done with animal anti-rhesus serum, it was found that more satisfactory results are obtained with human sera obtained from sensitized Rh-negative individuals. In fact, most of the antisera used today is derived by immunizing Rh-negative male donors with Rh-positive blood,[23] in order to stimulate the production of potent antibodies. It is not difficult to understand why human antisera give more satisfactory results than the anti-rhesus sera. The human antisera, having been prepared with human blood, are homospecific, while the anti-rhesus rabbit or guinea-pig sera are heterospecific. Therefore, the human antibodies fit the antigen precisely, like a key fits its own specific lock, while the anti-rhesus animal sera are in the nature of skeleton keys and react less avidly with the human Rh agglutinogen.[24]

Heredity of the Blood-Group Factors

The blood groups furnish the most striking example of Mendelian inheritance in man, and the recent work on the heredity of the Rh-Hr types, in particular, has stimulated a greatly increased interest in human

heredity. In the limited space available it will only be possible to outline these findings very briefly.

As has been demonstrated by Bernstein,[25] the four Landsteiner blood groups are inherited by a series of three allelic genes, most conveniently designated as I^o, I^A, and I^B, respectively, so that there are six genotypes corresponding to the four blood groups as shown in Table I. The dis-

TABLE 1. THE A-B-O PHENOTYPES
AND GENOTYPES

Phenotypes	Genotypes
O	$I^O I^O$
A	$I^A I^A$ and $I^A I^O$
B	$I^B I^B$ and $I^B I^O$
AB	$I^A I^B$

covery of variants of agglutinogen A, and the probable existence of similar variants of the agglutinogens B and O, indicate that the situation is actually far more complicated. If one takes into account only the major variants A_1 and A_2, a series of four allelic genes must be assumed.[26]

In Table 2 are given the reactions of the four agglutinogens corresponding to the four postulated genes.[27] The sera anti-A, anti-B, and anti-A_1 are the ones available for routine use; sera of specificity anti-O

TABLE 2. THE A-B-O SERIES OF GENES AND THE REACTIONS
THEY DETERMINE

Genes	Corresponding Agglutinogen	Reactions with Serums				
		Anti-A$_1$	Anti-A	Anti-B	Anti-C	Anti-O
I^{A_1}	A_1	+	+	−	+	−
I^B	B	−	−	+	+	−
i^{A_2}	A_2	−	+	−	−	+
i^O	O	−	−	−	−	+

are quite rare and are most often obtained from individuals of subgroup A_1B, while serum anti-C has only been encountered a few times and its reactions are not entirely certain.

According to the theory of Landsteiner and Levine [28] the three M-N types are inherited by a pair of allelic genes, preferably designated as L^x and L^y respectively, so that there are three genotypes corresponding to the three phenotypes. A priori one would expect that this theory over-simplifies the situation, as in the case of the theory of three allelic genes for the four blood groups. That the situation is far more complicated

has in fact been established by the discovery of the existence of the rare variants N_2 and M_2, and especially by the recent important discovery of Montgomery and Walsh[4] of a new agglutinogen, designated S by them, which is related to the M-N system of antigens. The agglutinogens M and N occur in at least two forms, one with, and one without agglutinogen S. Taking only the agglutinogens M, N, and S into account, it is necessary to postulate [29] a series of at least four allelic genes as shown in Table 3.

In the case of the Rh-Hr types,[30] there are available for general use Rh antisera of the three "standard" specificities, anti-Rh_o, anti-rh′ and anti-rh″, and in addition, a less common anti-Rh serum designated as anti-rh^w. Of three theoretically possible [31] Hr antisera only anti-hr′ and

TABLE 3. THE M-N SERIES OF ALLELIC GENES

Genes	Corresponding Agglutinogens	Reactions with Serums		
		Anti-M	Anti-N	Anti-S
L^M	M	+	−	−
L^{Ms}	Ms	+	−	+
L^N	N	−	+	−
L^{Ns}	Ns	−	+	+

anti-hr″ have been demonstrated beyond doubt.[32-34] According to the theory of Wiener,[30] the Rh-Hr blood types are inherited by a series multiple allelic genes, of which a minimum of eight "standard" genes has been proved [35] to exist, namely r, r', r'', r^y, R^0, R^1, R^2 and R^z. With the aid of the recently discovered anti-rh^w serum, a ninth allelic gene R^{1w} has been completely identified by Callender et al.,[36] who also state they have identified an additional agglutinogen giving reactions corresponding to those expected for gene r'^w. Wiener [27] has reported the existence of numerous bloods giving weak or intermediate reactions with Rh or Hr antisera, indicating the existence of numerous variants of the Rh-Hr factors, and this work has been confirmed by British workers.[37, 38] Therefore, the list of Rh series of genes given in Table 4 actually greatly over-simplifies the subject.

Knowledge of the inheritance of the Rh-Hr types is of value for predicting the outcome of pregnancies in which sensitized Rh-negative women have Rh-positive husbands.[39] Knowledge of the heredity of the blood groups in general has been applied in cases of disputed parentage. In such cases, obviously, the blood groups can be used only to exclude and not to prove parentage. The instances in which parentage can be excluded are summarized in the following laws[40]:

(1) The agglutinogens A, B, M, N, S, P, Rh_o, rh', rh", rh^w, hr', and hr" cannot appear in the blood of a child unless present in the blood of one or both parents.

(2) Group AB parents cannot have group O children, and group O parents cannot have group AB children.

(3) Type M parents cannot have type N children, and type N parents cannot have type M children.

(4a) rh'-negative parents cannot have hr'-negative children, and hr'-negative parents cannot have rh'-negative children.

(4b) rh"-negative parents cannot have hr"-negative children, and hr"-negative parents cannot have rh"-negative children.

If a man is falsely accused of parentage he can be exonerated, on the average, in about half the cases by the blood tests. The subgroups of A increases the chances only slightly, and are not considered sufficiently reliable for medicolegal cases. Blood tests have also been used in a number of cases where babies had been interchanged in hospitals. For further details concerning this aspect of the subject, the reader is referred to the excellent book by Schatkin.[41]

"Mosaic" Structure of Agglutinogens

Since, according to the modern concept, antibodies are formed in a manner similar to the preparation of a cast from a mold, there obviously is no one-to-one correspondence between antigens and antibodies, because a single antigen could give rise to a multiplicity of different antibodies. In a recent review Haurowitz [42] has presented a clear description of the manner in which antibodies are probably formed. First, the globulin molecule is reproduced in an extended form by the production of a positive replica of the peptide chain of a template.* The newly-formed globulin molecule then curls up to form an ellipsoid, and polar groups at the surface of the molecule adapt themselves to the surface of any foreign antigen contained in the antibody-forming cell so as to form a negative impression of the surface of the antigen.

Landsteiner [43] has indeed shown, with the aid of antigens containing simple hapten groups of known chemical composition, how a single antigen can give rise to a multiplicity of antibodies. For example, immune sera for m-aminobenzene sulfonic acid will cross-react with antigens containing o-aminobenzene sulfonic acid, m-aminobenzoic acid, and m-aminobenzene arsenic acid. By absorption experiments, it was possible to show that these cross-reactions were due not merely to a single poorly adapted antibody but to four distinct antibodies, and no doubt the number of antibodies demonstrable would have been increased had a larger number of cross-reacting antigens been tested.

* See page 243.

The principles discovered by Landsteiner by working with antigens of known chemical constitution must apply with equal force to antigens of unknown chemical structure. It is not remarkable therefore that red-cell agglutinogens, which probably have a molecular structure far more complicated than the chemicals tested by Landsteiner, behave in serological tests as if they had a "mosaic" structure. For example, by cross-testing anti-B sera with bloods of lower mammals, it was found [44] that the human B agglutinogen has a number of "components" or "partial antigens" designated as B_I, B_{II}, B_{III}. . . . Similarly, human A agglutinogen has so-called F_A components, as shown by tests with anti-sheep immune rabbit sera.[18] In addition A_2 and O have components in common which are absent from A_1 and B; A and B have components in common which are absent from O.[27] Moreover, by cross-testing anti-M and anti-N sera with bloods of apes and monkeys, at least five partial antigens have been demonstrated in human M agglutinogen, and at least two for human N agglutinogen.[17] Landsteiner's concepts have been applied when devising designations for the human agglutinogens. For example, based on their serological reactions, the subgroups of A were at first designated as AA_1 and AA_2.[45, 46] When, however, genetic studies proved that AA_1 and AA_2 were units transmitted by corresponding allelic genes, so that their reactions were due to partial antigens, the designations of the subgroups were simplified to A_1 and A_2, respectively.[47]

When the heredity of the A-B-O blood groups was first studied, three possibilities suggested themselves, namely, (1) inheritance of independent pairs of allelic genes, *A-a* and *B-b*,[48] (2) inheritance by linked gene pairs, (*Ab*), (*aB*), (*AB*), and (*ab*),[49] or (3) inheritance by multiple allelic genes, I^A, I^B, and I^O.[25] A fourth possibility, namely, completely linked gene couplets, is obviously indistinguishable from the case of multiple alleles, and it is mere quibbling to try to distinguish the case of completely linked genes from that of allelic unit genes.[50] Bernstein pointed out that if there were separate pairs of genes for A and B, then at equilibrium the distribution of the blood groups in a population would satisfy the relation, $AB \times O = A \times B$. However, with only rare exceptions, this relation does not hold, so that it became evident that the blood groups must be inherited by multiple allelic genes.

Similarly, in the case of the Rh-Hr types, one had to consider the possibility of four separate pairs of genes (either independent or linked) for Rh_o, rh^w, rh'-hr', and rh''-hr'', respectively, or the alternative possibility of multiple allelic genes. As shown by Wiener,[51] the distribution of the Rh types in the population does not conform with the expectations under the former hypothesis, but satisfies instead the theory of multiple allelic genes. It is clear, therefore, that the factors, Rh_o, rh', rh'', rh^w, hr', and hr'' do not represent separate agglutinogens determined by corre-

sponding genes, but instead partial antigens, as in the case of AA_1 and AA_2. Accordingly for agglutinogen Rh_1^w, for example, the simple designation Rh_1^w is more rational than $Rh_orh'rh^whr''$ (or CC^wDe in the British notations), besides having the important advantage of greater simplicity (cf. Table 4).[52]

In cattle, the existence of more than 40 antigenic differences has been demonstrated in the blood cells. In a recent important paper, Stormont[65] has shown that the resulting complex antigenic mosaic can be reduced to a few complex systems of agglutinogens, each transmitted by a series of multiple allelic genes, as in man.

TABLE 4. THE ALLELIC RH GENES AND THEIR CORRESPONDING AGGLUTINOGENS

Gene	Corresponding Agglutinogen	Reactions with Rh Sera				Reactions with Hr Sera		Partial Antigens Present in Agglutinogen
		Anti-Rh₀	Anti-rh'	Anti-rh''	Anti-rhʷ	Anti-hr'	Anti-hr''	
r	rh	−	−	−	−	+	+	hr' and hr''
r'	rh'	−	+	−	−	−	+	rh' and hr''
r''	rh''	−	−	+	−	+	−	rh'' and hr'
r^y	rh_y	−	+	+	−	−	−	rh' and rh''
R^o	Rh₀	+	−	−	−	+	+	Rh₀, hr', and hr''
R^1	Rh₁	+	+	−	−	−	+	Rh₀, rh', and hr''
R^2	Rh₂	+	−	+	−	+	−	Rh₀, rh'' and hr'
R^z	Rh_z	+	+	+	−	−	−	Rh₀, rh', and rh''
R^{1w}	Rh₁ʷ	+	+	−	+	−	+	Rh₀, rh', rhʷ, and hr''

Pathogenesis and Treatment of Erythroblastosis

According to the theory of Levine et al.,[22] in typical cases of erythroblastosis, an Rh-negative mother carrying an Rh-positive fetus becomes sensitized to the Rh factor, and the Rh agglutinins produced by her pass through the placenta and combine with the Rh-positive blood of the fetus, giving rise to one or another manifestation of erythroblastosis fetalis. This theory receives support from the observation that fully 90 per cent of mothers of erythroblastotic babies are Rh-negative in contrast to an incidence of only 15 per cent in the general population. However, it was difficult to explain why, in many of the most severe cases resulting in stillbirths, no Rh agglutinins were demonstrable in the maternal serum; the same problem also arose in some cases of intragroup hemolytic reactions caused by Rh sensitization.[53] As already mentioned, it was then shown that these hitherto obscure cases were due to univalent Rh antibodies. Based on the theoretic concepts, moreover, it was postulated that univalent antibodies are composed of smaller molecules than

bivalent antibodies, and direct tests did show that univalent Rh antibodies readily pass through the intact placenta while bivalent Rh antibodies are held back by the placental barrier. (Other differences between univalent antibodies and bivalent antibodies are shown in

TABLE 5. DIFFERENCES BETWEEN "UNIVALENT" AND "BIVALENT" ANTIBODIES

Characteristics	Bivalent Antibodies	Univalent Antibodies
Common names	Agglutinin; precipitin	Glutinin; blocking antibody
Usual time of appearance in course of immunization	Early	Late
Chemical nature	"Euglobulin"; precipitated by sodium sulfate solutions of concentrations 13.5 to 17.4 per cent	"Pseudoglobulin"; precipitated by sodium sulfate solutions of concentrations 17.4 to 21.5 per cent
Sedimentation constant	18	7
Probable molecular weight	930,000	155,000
Diffusibility	Does not diffuse readily	Diffuses readily
Resistance to heating	Relatively thermolabile	Relatively thermostabile
Reaction with cells in saline media	Clumps cells by agglutination	Coats cells without clumping them
Reaction with cells in plasma or serum media	Clumps cells by agglutination	Clumps cells by conglutination
Nature of clumps	Easily dislodged from glass surface	Tend to adhere to glass surface
Specificity of clumps	Specific — clumps contain only one type of cell	Nonspecific — clumps may contain more than one type of cell
Reaction with cells in presence of complement	Does not fix complement or lyse cells	Fixes complement and lyses cells
Behavior relative to placenta	Held back by the intact placenta	Passes through placenta readily
Role in erythroblastosis	Minor	Major
Role in immunity	Precipitating antibody	Protective antibody
Role in allergy	Sensitizing antibody (reagin)	Blocking antibody

Table 5.) Thus, it is clear that it is the univalent Rh antibodies (glutinins or blockers) which cause erythroblastosis and not the Rh agglutinins as originally postulated by Levine, who did have available knowledge concerning the existence of two sorts of Rh antibodies when he investigated the subject in 1940. In the exceptional 10 per cent of cases

where the mother is Rh-positive, isosensitization to the A-B blood factors, Rh-Hr subtypes, and other blood factors has been demonstrated.

That univalent Rh antibodies are the usual cause of erythroblastosis is proved by the correlation between the expectant Rh-negative mother's Rh antibody titer and the incidence of stillbirths.[55] In erythroblastotic babies born alive, the red cells are coated with univalent antibodies. Many such babies appear quite well at first, but as complement and
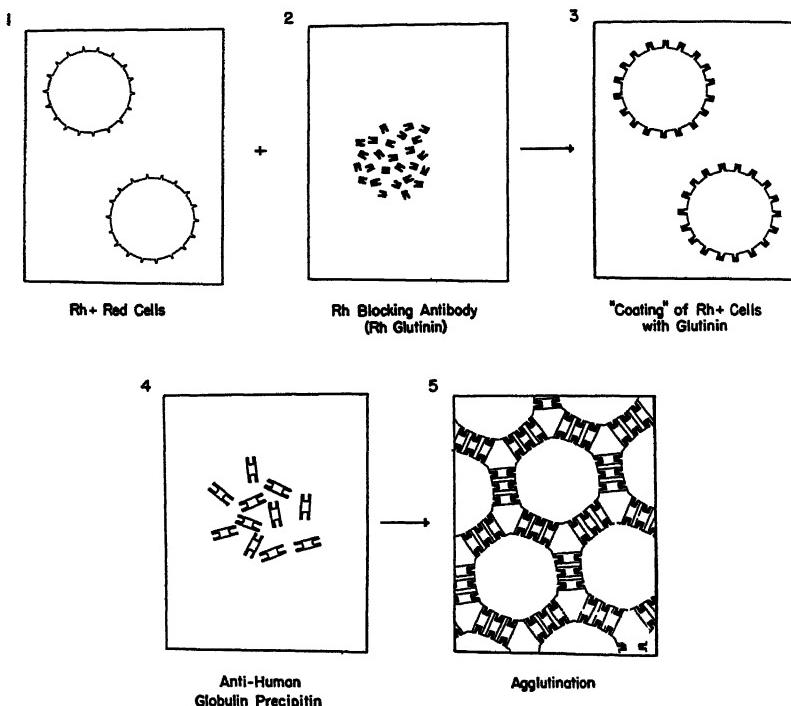
Figure 4. Diagrammatic representation of the antiglobulin test.

conglutinin mature in their sera, intravascular lysis and clumping of red cells cause the development of anemia and jaundice within a few hours or days after birth.[56] About 50 to 75 per cent of such untreated babies die, and among those who survive, about 10 to 20 per cent subsequently exhibit sequelae of marked mental retardation. By the operation of exchange transfusion, the mortality rate can be reduced to 10 per cent or less, and sequelae prevented. In this procedure, the baby's blood is withdrawn and simultaneously replaced by inagglutinable (Rh-negative) blood, the ideal procedure using about 1000 cc of blood (4 times a newborn baby's blood volume).[57]

For best results, the occurrence of erythroblastosis should be antici-
pated before the baby is born, and antenatal Rh typing of expectant
mother and father, and Rh antibody tests have become routine. To detect
and titrate univalent antibodies, the albumin-plasma conglutination
technic gives excellent results,[58] though other satisfactory methods have
also been devised, such as the antiglobulin test [59] (cf. Figure 4). The
blocking test is far too insensitive, because for blocking to occur all or
almost all the Rh haptens must be coated with antibody, while only a
small fraction of the Rh haptens need be coated for conglutination to
occur. Therefore, the direct conglutination test is 5 to 40 times as sensi-
tive as the blocking test. Nevertheless, the blocking test, the conglutina-
tion test, and the antiglobulin test all appear to detect the identical
antibody, and do not provide evidence for the existence of a third order
of antibodies in the sense suggested by certain workers. Recently Morton
and Pickles [60] have devised an ingenious test for univalent antibodies by
trypsinating the red cells; this new test gives the same results as the
conglutination and antiglobulin tests, but the exact mechanism by which
trypsinating red cells brings about clumping is still obscure. Certain high-
molecular substances such as gelatin, acacia, and dextran can function
as conglutinin substitutes in the conglutination test, but their use has
not found favor because of the tendency for nonspecific clumping.[61]

Concluding Remarks

In the limited space available, it has been possible only to outline
briefly the most important facts known concerning the human agglu-
tinogens. While only the most important allelic series, A-B-O, M-N, and
Rh-Hr were discussed, actually 7 separate series of allelic genes determin-
ing red-cell agglutinogens are known at present.[62] Owing to limitations
of space, no reference has been made to the important contribution
studies on the distribution of the blood factors have made to physical
anthropology.

While the specific agglutinins, like other antibodies, have been shown
to be serum globulins, little is known of the chemical nature of the red
cell agglutinogens and this will be an important subject for future in-
vestigations. The difficulty up to now has been to extract the red-cell
agglutinogens in pure form from the red cells, and in quantities sufficient
for chemical analysis. Some measure of success has been achieved with
group substances A and B because these occur in soluble form and in
large amounts in the secretions of a high percentage of individuals
(secretors), and substances with similar serological reactions are ubiq-
uitous in nature, occurring in many bacteria, and in animal secretions
such as horse saliva and hog pepsin. From pioneering work of Land-

steiner and Chase,[63] it appears that the A-B group substances are muco-
polysaccarides containing glucosamine and galactose, with amino-acids
probably an essential component. In recent work Kabat *et al.*[64] have
associated the A activity of preparations of hog and human origin with
their content of fucose.

The introduction of blood banks and increased safeguards surrounding
blood transfusion procedures have led to the increased use of this thera-
peutic measure, so that today blood is administered almost as freely as
saline. As a result, rare cases of intragroup hemolysis have come to light,
apparently due to isosensitization but without demonstrable *in vitro*
incompatibility. In addition, the intense interest focussed recently on
erythroblastosis fetalis has brought to light cases which are typical clin-
ically except for the absence of manifest serological incompatibility. Such
cases are possibly due to isosensitization against blood factors not de-
monstrable by techniques available at the present time. Perhaps such
factors occur too infrequently about the periphery of the cells to be
demonstrable by clumping or lysis, and special new methods may have
to be invented to demonstrate their presence. Eventually the day may
come when Landsteiner's prediction will be fulfilled, so that any person
can be identified by his blood almost as readily as by his fingerprints.

REFERENCES

1. Landsteiner, K., and Wiener, A. S., *Proc. Soc. Exptl. Biol. Med.*, **43**, 223 (1940).
2. ———, and ———, *J. Exptl. Med.*, **74**, 309-20 (1941).
3. Wiener, A. S., "Blood Groups and Transfusion," 3rd ed., 438 pp., Springfield, Ill.,
   Charles C. Thomas, Publisher, 1943.
4. Walsh, R. J., and Montgomery, C., *Nature*, **160**, 504 (1947).
5. Sanger, R., Race, R. R., Lawler, S. D., and Bertinshaw, D., *Heredity*, **3**, 205-13
   (1949).
6. Wiener, A. S., Proc. of Eighth International Congress of Genetics, *Hereditas*, Supp.
   Vol., 500-19 (1949).
7. Breinl, F., and Haurowitz, F., *Z. physiol. Chem.*, **192**, 45-7 (1930).
8. Alexander, J., *Protoplasma*, **14**, 296-306 (1931).
9. Mudd, S., *J. Immunol.*, **23**, 423-27 (1932).
10. Landsteiner, K., "The Specificity of Serological Reactions," rev. ed., Cambridge,
    Mass., Harvard Univ. Press, 1945.
11. Boyd, W. C., "Fundamentals of Immunology," 2nd ed., New York Interscience
    Publishers, Inc., 1947.
12. Pauling, L., and Campbell, D. H., *J. Exptl. Med.*, **76**, 211-20 (1942).
13. Wiener, A. S., and Karowe, H., *J. Immunol.*, **49**, 51-61 (1944).
14. ———, *Am. J. Clin. Path.*, **16**, 477-97 (1946).
15. ———, *Proc. Soc. Exptl. Med.*, **56**, 173-6 (1944).
15a. Race, R. R., *Nature*, **153**, 771-2 (1944).
16. Wiener, A. S., *J. Lab. Clin. Med.*, **30**, 662-7 (1945).
17. Landsteiner, K., and Wiener, A. S., *J. Immunol.*, **33**, 19-23 (1937).
18. Schiff, F., and Adelsberger, L., *Z. Immunitäts.*, **40**, 335 (1924).
19. Wiener, A. S., and Peters, H. R., *Ann. Internal Med.*, **13**, 2306-22 (1940).
20. ———, *Arch. Path.*, **32**, 227-50 (1941).
21. Levine, P., and Stetson, R. E., *J. Am. Med. Assoc.*, **113**, 126 (1939).

22. Levine, P., Katzin, E. M., and Burnham, L., *J. Am. Med. Assoc.*, **116**, 827 (1941); Levine, P., Burnham, L., Katzin, E. M., and Vogel, P., *Am. J. Obstet. Gynecol.*, **42**, 925-937 (1941).
23. Wiener, A. S., *Proc. Soc. Exptl. Biol. Med.*, **70**, 576-8 (1949).
24. ———, *Am. J. Clin. Path.*, **16**, 233-243 (1946).
25. Bernstein, F., *Klin. Wochschr.*, **3**, 1495 (1925); *Z. indukt. Abstamm. Vererbungsl.*, **37**, 237 (1925).
26. Landsteiner, K., and Levine, P., *Proc. Soc. Exptl. Biol. Med.*, **24**, 941-2 (1927); Thomsen, O., Freidenreich, V., and Worsaae, E., *Acta path. Microbiol. Scand.*, **7**, (1930).
27. Wiener, A. S., *Science*, **100**, 595-7 (1944).
28. Landsteiner, K., and Levine, P., *J. Exptl. Med.*, **47**, 757-75 (1928).
29. Sanger, R., Race, R. R., Walsh, R. J., and Montgomery, C., *Heredity*, **2**, 131-9 (1948).
30. Wiener, A. S., *Proc. Soc. Exptl. Biol. Med.*, **54**, 316-9 (1943); Wiener, A. S., and Landsteiner, K., *Proc. Soc. Exptl. Biol. Med.*, **53**, 167-70 (1943).
31. Wiener, A. S., *Science*, **102**, 479-82 (1945).
32. Levine, J., *J. Pediat.*, **23**, 656-75 (1943).
33. Mourant, A. E., *Nature*, **155**, 542 (1945).
34. Wiener, A. S., *Blood*, in press.
35. ———, and Hyman, M. A., *Am. J. Clin. Path.*, **18**, 921-6 (1948).
36. Callender, S. T., Race, R. R., and Paykoc, Z. V., *Brit. Med. J.*, **2**, 83-4 (1945); Wiener, A. S., and Gordon, E. B., *Am. J. Clin. Path.*, **19**, 621-9 (1949).
37. Race, R. R., Sanger, R., and Lawler, S. D., *Heredity*, **2**, 237-50 (1948).
38. ———, ———, and ———, *Ann. Eugenics*, **14**, 171-84 (1948).
39. Wiener, A. S., and Sonn, E. B., *Am. J. Diseases Children*, **71**, 25-40 (1946).
40. Wiener, A. S., *Bull. N. Y. Acad. Med.*, **25**, 255-60 (1949).
41. Schatkin, S. B., "Disputed Paternity Proceedings," 2nd ed., Albany, N. Y., Matthew Bender Co., 1947.
42. Haurowitz, F., *Quart. Rev. Biol.*, **24**, 93-101 (1949).
43. Landsteiner, K., and van der Scheer, J., *J. Exptl. Med.*, **63**, 325-39 (1936).
44. Friedenreich, V., *Z. Immunitäts.*, **78**, 152 (1933).
45. Landsteiner, K., Witt, D. H., *J. Immunol.*, **11**, 221-47 (1926).
46. ———, and Levine, P., *J. Immunol.*, **12**, 441-60 (1926).
47. ———, and ———, *J. Immunol.*, **17**, 1-28 (1929).
48. v. Dungern, E., and Hirschfeld, L., *Z. Immunitäts.*, **6**, 284 (1910).
49. Kirihara, and Haku, *Nagoya J. Med. Sci.*, **2**, 75 (1927); Bauer, K. H., *Klin. Wochschr.*, **7**, 1588 (1928).
50. Wiener, A. S., "Blood Groups and Transfusion," 3rd ed., pp., 185 and 195, Springfield, Ill., Charles C. Thomas, Publisher, 1943.
51. ———, "Blood Groups and Transfusion," 3rd ed., p. 254, Springfield, Ill., Charles C. Thomas, Publisher, 1943 [cf. Rife, D. C., *Ohio J. Sci.*, **48**, 116-8 (1948).]
52. ———, and Gordon, E. B., *Am. J. Clin. Path.*, **19**, 621-9 (1949).
53. ———, *Arch. Path.*, **32**, 227-50 (1941).
54. ———, *Ann. Allergy*, **6**, 293-305 (1948).
55. Sacks, M. S., Kuhns, W. J., and Jahn, S. F., *Am. J. Obstet. Gynecol.*, **54**, 400-14 (1947); Wiener, A. S., *Proc. Am. Assoc. Blood Banks*, Aug. (1948).
56. Wiener, A. S., *Exptl. Med. and Surg.*, **7**, 200-12 (1949).
57. ———, and Wexler, I. B., *J. Lab. Clin. Med.*, **31**, 1016-19 (1946); Wiener, A. S., and Wexler, I. B., *Blood*, **4**, 1-35 (1949); Wiener, A. S., *Postgrad. Med. J.*, **7**, 1-10 (Jan. 1950).
58. Wiener, A. S., and Hurst, J. G., *Exptl. Med. and Surg.*, **5**, 285-98 (1947).
59. Coombs, R. R. A., Mourant, A. E., and Race, R. R., *Brit. J. Exptl. Path.*, **26**, 255 (1945).
60. Morton, J. A., and Pickles, M. M., *Nature, London*, **159**, 779 (1947).
61. Levine, P., *Am. J. Clin. Path.*, **16**, 597 (1946); Fisk, R. T., and McGee, C. A., *Am. J. Clin. Path.*, **17**, 737 (1947).

62. Sanger, R., and Race, R. R., *Ann. Eugenics*, 15, 77-90 (1949).
63. Landsteiner, K., and Chase, M. W., *J. Exptl. Med.*, **63,** 813 (1936).
64. Kabat, E. A., Baer, H., and Knaub, V., *J. Exptl. Med.*, **89,** 1-10 (1949).
65. Stormont, C., *Genetics*, **35,** 76-94 (1950).

DENTAL CARIES

L. S. Fosdick

Professor of Chemistry, Northwestern University Dental School, Chicago, Ill.

DURING the past few decades we have seen dental research bring order out of apparent chaos, so that we now have an explanation for man} of the seemingly conflicting observations that have been made in the past.[1] The first major theory of dental caries formulated by Miller [2] and Black [3] indicated that the carious process was caused by microbial fermentation of carbohydrates with the formation of acids, particularly lactic acid. The substrate and the acid were supposedly held in place by a gelatinous plaque, so that the typical decalcified areas appeared in the teeth in the protected or unclean portions of the tooth surface. This theory did not provide any information concerning the methods or conditions under which the acids are formed nor did it explain many of the anomalies which are ever present in a condition dependent upon so many variables. According to the chemical parasitic theory, dental caries should be a function of the carbohydrate intake, but many clinical observations indicate that this is not necessarily true. Some inhabitants of various countries consume large quantities of carbohydrates yet have a low incidence of caries. On the other hand, there are individuals who consume small quantities of carbohydrates and have a relatively high incidence of dental caries. Furthermore, individuals with the same hereditary background, and even individuals in the same family, with similar diets, have different rates of caries formation. In view of the confusion that exists in interpreting the carious process in the light of our older theories and in view of the clarity with which the newer theories may present the problem, it would be well to examine all available data before conclusions are drawn.

Local Conditions for Carious Process

It is known that the carious process always starts from the surface of the tooth and penetrates toward the pulp and that this process is a

decalcification of the inorganic materials of the tooth, accompanied or
followed by a proteolysis of the organic materials. In the simplest form
one may consider a system consisting of vital teeth which are normally
bathed in saliva and are subjected to all the forces that exist in the
mouth. Under these conditions, if the chemistry of the tooth substance
is understood, and the composition and reaction of the saliva are
known, we should be able to determine the variables affecting the
tooth substance.

Tooth Structure. An analysis of the tooth indicates that it is essen-
tially a calcified organic material. The pulp on the inside of the tooth
has an adequate blood supply containing all the ions and ingredients
normally found in blood. Surrounding this is the dentine and cemen-
tum, which so far as is known have no blood supply. It does contain
considerable quantities of moisture. The moisture may or may not exist
in definite channels; there is still considerable difference of opinion
concerning the channels for the dental lymph. It is known, however,

TABLE 1. WATER, ORGANIC AND INORGANIC MATERIAL
IN THE MOUTH

	Enamel	Dentine	Cementum
H_2O	2.3	13.2	20
Organic	1.7	17.5	30
Inorganic	95.6	69.3	50

that the dentine and cementum are highly calcified, having about 75
per cent inorganic material. The enamel contains only traces of mois-
ture and organic material. In fact, the only evidence of organic ma-
terial in the intact enamel is found in the so-called lamellae. By means
of proper decalcification, a slight amount of organic material may be
demonstrated in the enamel, but the fact that it has not been shown
to exist without previous decalcification indicates that even the major
portion of this small amount of organic material may under normal
conditions be in the form of calcium salts, probably the calcium salt of
a protein (Table 1).

It has been conclusively shown that the inorganic portion of the
bone, dentine, cementum, and enamel is essentially the same, a major
difference between these structures being the amount of inorganic
material, rather than the kind of inorganic substances present
(Table 2). This inorganic phase of bone and teeth has been ex-
tensively studied both from a chemical and crystallographic point of
view. It has been shown that the inorganic crystalline substance is
definitely not a pure compound. It consists primarily of a contam-
inated calcium phosphate. Whether the calcium phosphate crystals are

pure calcium phosphate having impurities in solid solution, or whether the ions other than the calcium phosphate exist in a regular oriented crystal pattern is not known. Some investigators have assigned the formula, $Ca_3(PO_4)_2 \cdot nCaCO_3$, while others claim that the substance more nearly resembles apatite, $Ca_{10}(PO_4)_6(OH)_2$. Insofar as apatite, a hydrated calcium phosphate, is an ionic crystalline compound, it lends itself readily to x-ray analysis for crystal pattern. This has been exhaustively studied by several investigators. The first investigators, Klement and

TABLE 2. COMPOSITION OF INORGANIC PORTION
OF THE TEETH

	(%)
Ca—35.7 (as CaO)	50.0
P—17.0 (as P_2O_5	39.0
CO_2	4.0
H_2O	2.0
Mg—0.9 (as MgO)	1.5
Na—0.7 (as Na_2O)	1.0
K—0.16 (as K_2O)	0.2
Cl	0.05
Pb	0.02
Fe	0.02
F	0.02
Total	97.81

Trömel,[5,6] came to the conclusion that the crystalline material was relatively pure hydroxy apatite with other substances, such as magnesium phosphate, magnesium carbonate and calcium carbonate in solution and not in the regular crystalline pattern. These investigators found that if powdered bone or enamel were washed well with distilled water, the impurities could be removed, leaving relatively pure hydroxy apatite. The crystal pattern obtained by x-ray-diffraction analysis was identical with that obtained by mineral apatite. Bale, Le Fevre, and Hodge[7] later came to similar conclusions by means of x-ray-diffraction analysis.

Whether or not the carbonate and/or magnesium ions take part in the crystalline pattern or simply in solid solution is immaterial, as tooth substance should follow the ordinary laws of solubility.

The Chemical Behavior of Calcium Phosphates. Perhaps an investigation of the behavior of the various types of calcium phosphates would explain some of the behavior of tooth enamel. In 1923 Wendt and Clark[8] made a comprehensive study of the system: calcium hydroxyde plus phosphoric acid.

$$Ca(OH)_2 + 2H_3PO_4 \longrightarrow CaH_4(PO_4)_2 + 2H_2O$$

$$Ca(OH)_2 + H_3PO_4 \longrightarrow CaHPO_4 + 2H_2O$$

$$3Ca(OH)_2 + 2H_3PO_4 \longrightarrow Ca_3(PO_4)_2 + 6H_2O$$

$$10Ca(OH)_2 + 6H_3PO_4 \longrightarrow Ca_{10}(PO_4)_6(OH)_2 + 18H_2O$$

From their studies it is obvious that the only stable form of calcium phosphate above pH 5.2 is a hydrated calcium phosphate or apatite. Other research workers (Holt, LaMer, and Chown,[9] and Fosdick and Starke [10] have indicated that the inorganic portion of teeth and bone is very similar to the only stable form of calcium phosphate above pH of 5.2. The solubility products were similar and the solubility curves
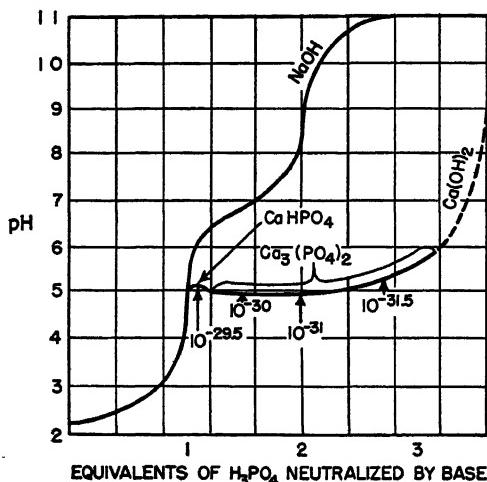
Figure 1. Titration curves of H_3PO_4 with NaOH and $Ca(OH)_2$.

followed the same general trend as those determined by Wendt and Clark, using pure chemicals.

Recently, Kuyper [11] made a comprehensive study similar to that of Wendt and Clark; however, instead of pure solutions he used solutions similar in composition to blood plasma for the precipitating media. He found that the resulting precipitate was identical in composition to that of bone and tooth salts. Thus, a very common chemical phenomenon is displayed. When insoluble salts are precipitated from solution, in a number of cases, depending to a large extent upon the salts present, many ions of different solubilities are "carried down" with the precipitates. This is especially true with calcium and barium salts. When one considers the mechanism whereby bone salts are precipitated to form bone and the hard structures of the body, we find a condition very similar to that which existed in the experiment of Kuyper.

One may write the equation for the solubility product of calcium phosphate as follows:

$$[Ca^{++}]^3 \times [PO_4^{=}]^2 = K = 10^{-27}$$

It is admitted that tricalcium phosphate is not the same as the hydrated form but the solubility curves should be very similar, as shown by the

work of Holt, LaMer, and Chown. In this equation the calcium ion, cubed, times the tertiary phosphate ion, squared, is equal to a constant— the solubility product. The blood plasma and interstitial fluids are normally saturated or supersaturated with calcium and phosphate ions. Hence any local increase in either the calcium or phosphate ion will cause a deposition of solid calcium phosphate, and if the pH is above 5.2 the calcium phosphate precipitated will be in the hydrated form.

It has been quite adequately demonstrated that the mechanism whereby bone salts are laid down in the organic materials is due to a local increase in phosphate ion.[12] It has been shown that the organic phosphates in the blood, which are normally un-ionized, are hydrolyzed to the free phosphate ion under the influence of certain phosphatase enzymes. It has also been shown that the osteoblasts of the bone or tooth organs contain a large amount of the enzyme, phosphatase, and hence when circulating blood or lymph contacts these bone-forming organs, there is a localized increase in phosphate ions, thus causing the solubility product to be exceeded, with the subsequent precipitation of the solid crystalline material. These organs, through some unknown mechanism, cause an orientation of the crystals so that the normal pattern for bone, cementum, dentine, and enamel are formed.* On the basis of the above, it is obvious that the major constituent of the inorganic phase of tooth enamel is a hydrated calcium phosphate, and as the precipitation of bone salts follows the usual rules of solubility, it would be expected that the inorganic crystals in enamel and bone would be a contaminated calcium phosphate rather than the pure compound. The contamination would be comprised of those ions which are normally carried down when calcium phosphate is precipitated from blood plasma or lymph. Thus, those factors influencing the solubility of tooth enamel would be the same factors that would influence the solubility of tricalcium phosphate, with the modifications that would result from contamination with other ions found in blood plasma.

Insofar as experimental evidence has indicated that the solubility of human enamel is slightly greater than the solubility of tricalcium phosphate, we would expect that the enamel would dissolve under less rigorous conditions than if it were a pure compound. On the basis of the solubility product, however, it is quite clear that the two factors that would influence solubility to the greatest extent would be an alteration of the calcium-ion concentration and/or an alteration of the phosphate-ion concentration. There are many substances that would cause a decrease in the calcium-ion concentration. Such substances as

* In this connection see a much neglected book by William M. Ord (London, 1879), entitled "The Influence of Colloids on Crystalline Form and Cohesion," which discusses bone formation and the action of protective colloids in the light of then-existing information. In the human organism protective colloids are ubiquitous.—*Ed.*

citrates, lactates, tartrates, or even sugars would bring about the formation of un-ionized calcium complexes, thus causing a decrease in the calcium-ion concentration. Other substances such as the hexametaphosphates and some of the organic ion-exchange compounds may do the same thing. Thus, it is possible that these substances may, if found in saliva or in the mouth, cause a depletion of the calcium-ion concentration with the subsequent decalcification of the enamel. The only such substances that are likely to be found in the mouth are lactates, citrates, and sugars. None of these, with the possible exception of lactates, materially occurs in the saliva, but citrates and sugars may be present for short periods of time as a result of the ingestion of foods containing these substances. The effect of these compounds on the solubility of calcium phosphates should be of little importance because the saliva always contains an excess of calcium ions. However, these factors in regard to the calcium ion cannot be ignored, because of the varying local conditions found in the mouth.

There are few factors that will alter or influence the phosphate-ion concentration. Some of the anion-exchange substances which recently have been prepared could decrease the phosphate concentration, but most likely a change in the hydrogen-ion concentration would bring it about. The effect of a hydrogen ion is very profound and under certain conditions may cause a total depletion of the tertiary phosphate ion. This can best be explained by the following equations:

$$H_3PO_4 \rightleftharpoons H^+ + H_2PO_4^-$$
$$H_2PO_4^- \rightleftharpoons H^+ + HPO_4^=$$
$$HPO_4^= \rightleftharpoons H^+ + PO_4^\equiv$$
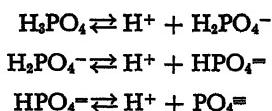
On the basis of the above, it is obvious that with an increase in hydrogen-ion concentration the reaction tends to go to the left, thus causing a decrease in the tertiary phosphate ion with a subsequent increase in the secondary and primary phosphate-ion concentrations and un-ionized phosphoric acid. It has been rather conclusively shown that at a pH of 5.2 or below, the tertiary phosphate ion is depressed to zero, and hence on the basis of the laws of solubility, the solubility product cannot be satisfied. Thus, at a pH of 5.2 or below, tricalcium phosphate or the hydrated forms cannot exist, and all of the phosphates are transferred to the secondary or primary phosphates. Under these conditions the rate of solution would depend primarily upon the surface exposed. Fortunately this is never large.

Thus, on the basis of the composition of the tooth, the only factors that would influence the solubility under conditions existing in the mouth are the increase in hydrogen-ion concentration or the presence

of lactates, citrates or sugars. This has been the subject of many excellent investigations. Benedict and Kanthak [13] studied the effect of hydrogen-ion concentration on human tooth enamel in buffered solutions. Although the buffers used contained no calcium or phosphate ions and hence were not comparable to saliva, the results of their work were in accord with the theoretical aspects of the solubility of the calcium phosphates. In the same year Enright, Friesell, and Trescher,[14] in a very comprehensive study, among other things determined the hydrogen-ion concentration at which decalcification was initiated. Here again a theory was confirmed. It was found that little or no decalcification would occur above pH values of 5.2. In the following year Forbes [15] also studied this system, with comparable results. In 1938, Fosdick and Starke [10] found that the solubility of enamel in human saliva substantiated, in part, the theoretical aspects of the problem, but that the actual solubility curve was slightly higher than the theoretical. Thus, the enamel was found to be slightly more soluble in saliva than it should be, and it was found that the solubility was a function not only of the hydrogen-ion concentration, but also of the calcium and phosphate normally present in the saliva.

The Chemistry of Saliva. Insofar as the saliva must be the carrier for any ionic solution of the tooth, the composition and variation in composition of the saliva might indicate the cause of the destruction of the inorganic portion of the tooth enamel. The saliva as secreted is a water solution containing about 0.5 per cent solids. These solids are composed of the following constituents:

Name	Amount (mg/100 cc)	Name	Amount (mg/100 cc)
Total solids	386–860	Nonprotein nitrogen	5.6–25.7
Calcium	4–10	Ammonia nitrogen	2.1–13.2
Phosphorus (inorganic)	10–30	Urea nitrogen	0–6.7
Chloride	30–60	Uric acid	0.5–2.9
Carbonate	55–125	Ptyalin	
Magnesium	1.29	Catalase	
Thiocyanate	30.90	Maltase	
Potassium	55.89	Urease	
Protein	200–400	Protease	
Mucin	0.25	Phosphatase	

The composition of saliva has been exhaustively studied. Perhaps the most thorough study is that of Becks and co-workers.[16–23] These men determined the calcium and phosphorus content of the saliva under practically all conceivable conditions. They found that the composition of saliva was variable, the concentration of its elements differing from individual to individual, and under various conditions in the same individual. Furthermore, the rate of flow of saliva was another non-constant factor. Of all the inorganic constituents of the saliva, only the calcium and phosphorus, carbonate and magnesium would have an ap-

preciable tendency to regulate the solubility of the enamel. The mag-
nesium and the carbonates would depress the solution of magnesium
carbonate and calcium carbonate, which are presumably impurities dis-
solved in the hydroxyapatite of the enamel. The calcium and phos-
phorus would tend to furnish the ions so that the solubility product
of the hydroxy apatite would be satisfied. Of course, it is understood
that the total calcium and the total phosphorus may play very little part
in producing a saturated solution of these ions. The important part of
the calcium or phosphorus is the calcium ion or the tertiary phosphorus
ion. Unfortunately there has been little work done concerning the
un-ionized phosphorus and calcium. The calcium-ion concentration is
dependent upon the protein complexes in the saliva and also upon the
presence or absence of such chemicals as citrates, lactates, and sugars.
Fortunately, free sugar has never been demonstrated in the saliva,
although reducing substances are present and galactose is present as an
integral part of the mucin molecule. The tertiary phosphate-ion con-
centration would depend primarily upon the hydrogen-ion concentra-
tion of the saliva. Fortunately, the hydrogen-ion concentration of saliva
has been exhaustively studied.[24-31]

It has been established that the normal hydrogen-ion concentration of
the saliva is around pH 6.8 during the basal flow of saliva. Although
there is considerable deviation from this value for different individuals,
the pH of the saliva has never been recorded below pH 6, by means
of accurate recording instruments. When the saliva is stimulated, as
occurs with the ingestion of any food or during mastication, the pH of
the saliva invariably rises and in many cases exceeds pH 8.4. Thus, it
would seem that a sufficient supply of the tertiary phosphate ion is
present in all freshly secreted saliva. The degree of ionization of these
constituents has never been determined, but preliminary work in the
laboratory of Northwestern University Dental School indicates that the
calcium is affected more than the phosphorus. The nondiffusible calcium
and phosphorus seem to be a function of erosion more than of dental
caries (Table 3).[32]

TABLE 3. NONDIFFUSIBLE Ca AND PO₄

	Erosion (%)	Nonerosion (%)
PO₄	25	16
Ca	27	16

Furthermore, it would seem that the offending material may be citrates,
which are present in erosion saliva in higher concentrations than in
either normal caries-free or caries-active saliva (Table 4).

Although the calcium and tertiary phosphate-ion concentration of
saliva has never been determined, it is known that freshly secreted saliva

is always supersaturated with these ions, except possibly in the case of individuals suffering from erosion. This fact can readily be confirmed by shaking sugar-free saliva with enamel particles. The results of work

TABLE 4. CITRIC ACID CONTENT IN SALIVA

Erosion Cases:

Case No.	Citric Acid in Saliva (mg/100 cc)	Case No.	Citric Acid in Saliva (mg/100 cc)
1	0.368	17	0.256
2	0.264	18	0.648
3	0.368	19	0.056
4	0.184	20	0.080
5	0.288	21	0.168
6	0.256	22	0.208
7	0.150	23	0.264
8	0.376	24	0.040
9	0.200	25	0.528
10	0.185	26	0.324
11	0.416	27	0.220
12	0.416	28	0.228
13	0.136	29	0.172
14	0.144	30	0.220
15	0.288	31	0.228
16	0.200	*Average*	0.254

Nonerosion Cases:

Case No.	Citric Acid in Saliva (mg/100 cc)	Case No.	Citric Acid in Saliva (mg/100 cc)
1	0.176	8	0.128
2	0.208	9	0.172
3	0.224	10	0.296
4	0.208	11	0.136
5	0.208	12	0.200
6	0.176	13	0.192
7	0.200	*Average*	0.194

of this nature indicate that all salivas have a different degree of super-saturation, and that variation occurs from individual to individual (Table 5).[33]

TABLE 5. SATURATION VALUE OF SALIVAS FOR ENAMEL
(12 cases of each)

Saliva Type	Ca	pH
Caries-free	6.1–3.0	7.0–7.2
Caries-susceptible	6.0–4.4	6.9–7.0
Erosion	6.2–5.8	7.0–7.1

Karshan[34] has shown that those individuals suffering from rampant caries have saliva that is not as saturated in respect to these ions as is the saliva from immune individuals. Whether or not this is due to a depression of the calcium ions or phosphate ions is not known, but in any event, calcium phosphate will invariably precipitate, except perhaps for erosion saliva.

On the basis of the above, it is obvious that the saliva as secreted has a powerful protective action against decalcification. Owing to the super-saturated state of this fluid a considerable depression of the calcium and/or phosphorus may take place without affecting the tooth structure, so that unless there are other agencies or factors that would locally reduce the calcium and/or phosphorus concentration, one could cate-gorically say that the teeth would not develop dental caries. However, there are many local factors which could, under certain conditions, depress calcium and phosphorus ions in particular areas of the mouth,
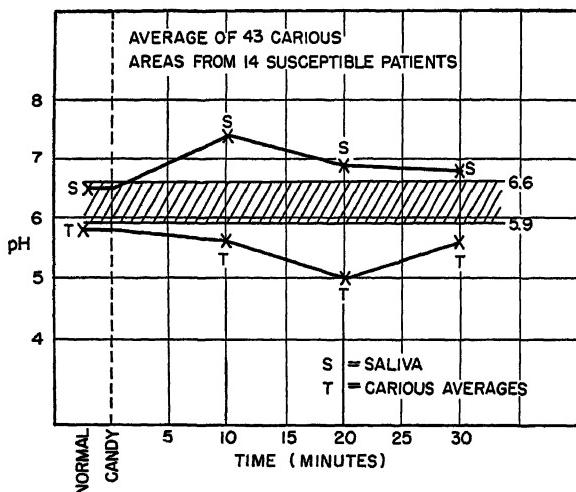
Figure 2. Changes in pH of saliva of carious areas following application of sugar.

thus causing many selective regions of decalcification. Any surface of the tooth that is not exposed to a continuous flow of freshly secreted saliva would be subjected to these local conditions, or any portions of the tooth in which food had lodged (such food becoming fermented or impacted without being washed out by the saliva and forces of mas-tication) could readily be subject to decay. It was one of the primary contentions of Black that carious lesions appear in well-defined areas be-cause of the protective action of impacted foods and the mucin plaque that is deposited in the so-called unclean areas of the mouth. Miller suggested that the impacted foods would slowly ferment, thus causing increased acidity in the localized areas. It was not considered that a very slow formation of acid would be likely to become neutralized by the slow seepage of saliva into these areas. The presence of a high hydrogen-ion concentration in these localized areas was suspected, but never actually demonstrated until Etherington and Trimble,[35] in 1934,

demonstrated pH values ranging from 4.6 to 6.8 in eighteen individuals. The pH of the interproximal spaces of caries-active and caries-immune cases averaged about the same.

In 1938, Stephan [36] measured the pH of 211 samples of plaque material and found that it ranged from 4.6 to 7.0. The average pH was 5.9, which is considerably above that required for decalcification of the teeth. From this report it would seem that the acidity of a carious area is not ordinarily that which would cause a decalcification but that occasionally a hydrogen-ion concentration sufficient to cause decalcification is observed. In 1940 Stephan,[37] by the use of an antimony electrode, found that the pH of carious areas would decrease to 4.5 in as short a time as two minutes, after the application of a ten per cent glucose solution, thus, for the first time indicating that there is a rapid conversion of sugars to acids. Fosdick, Campaigne, and Fancher [38] demonstrated the same effect by using a glass electrode and found that in many cases the application of sugar would cause an increase in hydrogen-ion concentration to pH 4 in as short a time as three minutes. In most cases, however, the drop in pH was less and the maximum acidity was twenty minutes after sugar was eaten (Figure 2).

Fancher and Fosdick [39] demonstrated that concurrently with a decrease in hydrogen-ion concentration there was an increase in lactate which varied from fifty to two hundred per cent. Furthermore, it was demonstrated that the high acid potential may remain for a period of from thirty to ninety minutes after the application of a fermentable carbohydrate. From this work it would seem that local increases in acidity can and do exist and that ordinary factors in the mouth will cause a neutralization of this acid within a very short time. The maximum acid potential is on the average reached in eighteen to twenty minutes, after which there is a gradual neutralization to a slightly lower pH value than that of the saliva.

Production of Acid from Sugar

Although it was not until 1940 that the production of acid from sugars was actually demonstrated in the oral cavity, much information concerning the mechanism of acid formation was available as early as 1935. At that time it was evident that the mechanism whereby sugars are converted to acids is a universal phenomenon and occurs not only under the influence of muscle and plant enzymes, but also under the influence of microbial enzymes. The sequence of the discoveries leading to our present-day concept of the phenomenon is extremely interesting and has largely been developed in the last fifty years.

Early Discoveries on Fermentation of Sugars. Fermentation of sugars is as old as recorded history, and no date can be assigned to the first observations of this phenomenon. Among the first to observe that the process of fermentation may be due to or associated with living organisms were Cagniard, Schwann, and Kützing.[40-42] These men based their observations purely on microscopic evidence, which in their day was not considered the highest type of evidence. The idea that fermentation was part of a living process was ill-received by scientists, and Berzelius, the leading chemist of the day, reviewed the articles with scorn. He considered that the process of fermentation was purely catalytic, the yeast organism being the catalyst, just as aluminum oxide, which had been demonstrated as possessing catalytic activity in some respects.[43] Wöhler and Liebig[44] severely criticized the authors, and even prepared an elaborate skit ridiculing Cagniard, Schwann, and Kützing. They described the yeast organism as little eggs that hatched into small animals in the shape of distilling apparatus. These animals were fed sugar in the retort end of the apparatus; alcohol and carbon dioxide were distilled.

The bitter controversies concerning the mechanism of fermentation lasted for several years until Pasteur[45] definitely proved the association of fermentation with the living yeast particles. The experiments of Pasteur were so convincing that even Liebig and Wöhler were forced to withdraw from the controversy.

Although Pasteur's concepts of the association of fermentation with life were accepted at the time, the chemists were still curious about the method whereby yeast caused the reactions. Traube[46] proposed that fermentation was due to substances similar to pepsin or enzymes liberated by the living organism. When Berthelot[47] isolated invertase from yeast, the theory of Traube gained limited acceptance. Berthelot, a famous French chemist, at the risk of gaining the displeasure of Pasteur, definitely expressed the opinion that insoluble ferments were present in the tissue of the yeast organism, and that the organism served primarily as an enzyme factory. He suggested that the only function of the yeast was to form enzymes which would, in turn, cause the chemical reactions to proceed. Berthelot, however, had no good experimental evidence to substantiate the theory, and Pasteur's contentions remained unchallenged until 1897. At this time Buchner,[48] a Dutch chemist, provided indisputable evidence that Berthelot's theory was correct. He extracted the enzymes from the yeast organism which, in the presence of phosphates, would cause fermentation of sugar with the formation of alcohol and carbon dioxide—entirely in the absence of living organisms. With the evidence provided by Buchner, it was obvious that both sides of the controversy were partly correct. Fermentation was associated with the living process of the yeast organism, but only insofar as it manufac-

tured the catalysts or enzymes necessary for the chemical reactions to take place.

With the information and experimental methods made available by Buchner's experiments, fermentation reactions were thoroughly studied and in the next three decades a rather complete picture of the reactions that occurred during the fermentation of sugar to alcohol and carbon dioxide was ascertained. During the early years of the investigations the yeast organism was used as the source of enzymes and coenzymes. Concurrently with the investigation of the enzyme systems of yeast, there was much interest in the biochemistry of muscle, muscle metabolism and muscle contraction. In the latter classical researches it was found that the same intermediate products were formed by the breakdown of glycogen during muscle contraction and by the metabolic utilization of sugar in the process of metabolism. Consequently, the investigators of fermentation and of muscle metabolism joined forces; the resulting progress was fruitful and rapid. It is not within the scope of this paper to discuss in detail the chemical research leading to the formulation of the exact chemical reactions that take place during fermentation. Suffice it to say that the researches of Harden, Young, Robinson, Embden, Meyerhof, Neuberg, Lohmann, Nielson, Euler, Werkman and Virtanen played important roles in solving the problem.[49-80]

Present-Day Views on Chemistry of Carbohydrate Metabolism. Carbohydrate metabolism in both plants and animals follows the same general pattern, indicating that the successive steps are mediated by similar enzymic catalysts, at least up to the final stages. Thus, when acidogenic bacteria act on sugar, the initial reaction stages resemble those found in the metabolism of carbohydrates by animal tissue; but in the final stages specific enzymes may determine the emergence of the end product, e.g., alcohol, acetic, propionic, butyric, lactic acid, etc. During the germination of seeds, carbohydrates are utilized for the production of energy by the same reactions. Present-day views of the mechanism are given in the following paragraphs.

(1) Phosphorylation of Glucose. The first reaction in the phosphorylation of glucose is the formation of glucose-6-phosphate under the influence of the enzyme, hexokinase:

$$\textit{Glucose} \underset{\substack{\text{myokinase, hexokinase,}\\\text{and ADP}}}{\overset{\substack{\text{hexokinase and ATP}}}{\rightleftharpoons}} \textit{Glucose-6-phosphate}$$
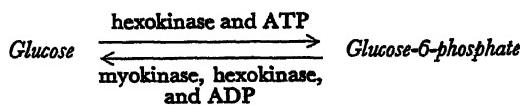
The phosphate is derived from adenyltriphosphate (ATP). This compound is a high-energy phosphonucleotide and has been isolated from yeast and various types of muscle tissue.[81-84] The enzyme, hexokinase, is a protein and occurs abundantly in yeast; it has been demonstrated in

other microorganisms. This enzyme catalyzes the reaction between ATP and glucose, with the formation of glucose-6-phosphate.[85]

In muscle tissue the reaction is stimulated to a considerable degree by myokinase, which causes the utilization of phosphate from the lower-energy-level compound, adenyldiphosphate, a phosphonucleotide found abundantly in muscle tissue. This nucleotide is usually designated as ADP. Myokinase,[85] a protein, is a rather stable enzyme. It is more stable at higher temperatures than is hexokinase, and is also stable toward acids. It is deactivated with hydrogen peroxide but may be reactivated with cystine or glutathione. The mechanism of myokinase action has been investigated by Kalckar, who found that this enzyme converts ADP into ATP. For this reason the reaction in the final analysis is essentially the same. Myokinase has never been demonstrated in microorganisms.

(2) Formation of the Neuberg Ester—Fructose-6-Phosphate. After the glucose-6-phosphate is formed, it is immediately acted upon by the enzyme system, consisting of phosphohexose isomerase and a metal, magnesium. This enzyme, a protein, is present in yeast, muscle tissue and a number of acidogenic bacteria. The enzyme, in the presence of magnesium ions, hastens the equilibrium between glucose-6-phosphate and fructose-6-phosphate.

$$\textit{Glucose-6-phosphate} \underset{\text{Mg}}{\overset{\text{phosphohexose isomerase}}{\rightleftarrows}} \textit{Fructose-6-phosphate}$$

Glucose-6-phosphate   *Fructose-6-phosphate*
(Robinson ester)   (Neuberg ester)

[Embden ester = 70%Robinson ester + 30% Neuberg ester]
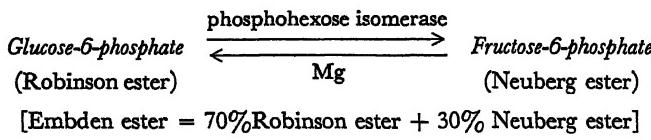
During the early years of investigation there was considerable confusion concerning this reaction. Embden isolated what he thought was a sugar phosphate from the reaction mixtures in which he had used tissue enzymes. Later evidence secured by Neuberg indicated that the substance isolated by Embden was in reality two sugar phosphates, glucose-6-phosphate and fructose-6-phosphate. The equilibrium has been thoroughly studied and found to consist of 70 per cent glucose-6-phosphate and 30 per cent fructose-6-phosphate.

(3) Formation of Fructose-1,6-Diphosphate (Harden Young Ester). Although the equilibrium between glucose-6-phosphate and fructose-6-phosphate is 70 per cent in favor of the synthesis of glucose, the reaction proceeds to the formation of lactic acid. As fast as the fructose-6-phosphate is formed, another enzyme, phosphohexokinase, causes the reaction to proceed to fructose-1,6-diphosphate.

$$\textit{Fructose-6-phosphate} \underset{\text{ATP}}{\overset{\text{phosphohexokinase}}{\rightleftarrows}} \textit{Fructose-1,6-diphosphate} + \textit{ADP}$$

Fructose-6-phosphate   *Fructose-1,6-diphosphate + ADP*
(Neuberg ester)   (Harden-Young ester)
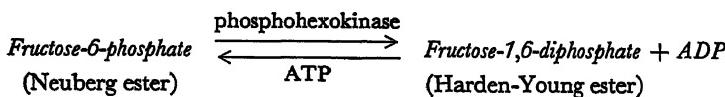
The enzyme has been isolated from yeast and from muscle tissue and has been shown to exist in a large number of sugar-consuming micro-organisms. The enzyme is a water-soluble protein. Here again ATP acts as a phosphate donor. The reaction yields the hexosediphosphate and ADP.[87-88]

(4) Dismutation of Fructose-1,6-Diphosphate. At this point in the series of reactions a dismutation of the hexose phosphate takes place to form two portions, each consisting of a molecule containing three car-bons. This reaction was one of the later problems solved, primarily because the two portions formed by the dismutation are in equilibrium.

$$H—C\!=\!O$$
$$H—C—OH \qquad \textit{Glyceric aldehyde}$$
$$H—C—OPO_3H_2 \qquad \textit{phosphate}$$
$$H$$

Fructose-1,6-diphosphate $\xrightleftharpoons[\text{}]{\text{zymohexase}}$ phosphotriose isomerase
(Harden-Young ester)

$$H—C—OH$$
$$C\!=\!O \qquad \textit{Dihydroxyacetone}$$
$$H—C—OPO_3H_2 \qquad \textit{phosphate}$$
$$H$$
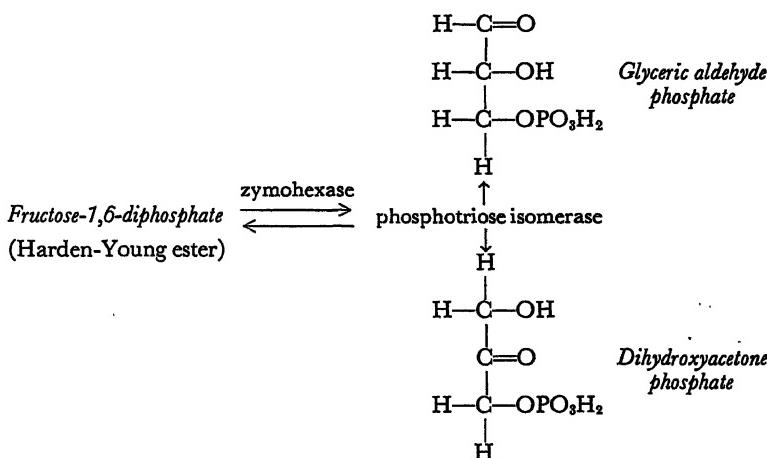
In the above, the reaction takes place under the influence of zymo-hexase. This enzyme has been isolated from muscle tissue and yeast.[89,90] It is a protein that is very easily inactivated by metal-binding sub-stances such as the various pyrophosphates, cystine and glutathione. It occurs in very high concentration in rat muscle, 1460 gamma per gram.

The equilibrium mixture of triosephosphates formed in this reaction has been intensively studied. The enzyme, phosphotriose isomerase, is a protein very abundant in muscle tissue and yeast. It establishes an equilibrium in favor of the dihydroxyacetone phosphate, which is a nonreactive compound, so that, the reaction proceeds through the gly-ceric aldehyde phosphate. The enzyme requires Mg for its activation.

(5) Formation of 1,3-Diphosphoglyceric Aldehyde. The next step in the production of lactic acid is the nonenzymic formation of 1,3-diphos-phoglyceric aldehyde. This reaction has not as yet been definitely proven. The compound has not been isolated except in the form of the dimer and its existence is postulated only on the basis of certain physi-cal characteristics of the reaction mixture.

$$
\begin{array}{ccccc}
\text{H} & & \text{H} & & \text{OH} \\
| & & | & & | \\
\text{H---C---OH} & & \text{C=O} & & \text{H---C---OPO}_3\text{H}_2 \\
| & \text{phosphotriose} & | & \text{nonenzymic} & | \\
\text{C=O} & \text{isomerase} & \text{H---C---OH} & \xrightleftharpoons{} & \text{H---C---OH} \\
| & \xrightleftharpoons{} & | & \text{PO}_4^- & | \\
\text{H---C---OPO}_3\text{H}_2 & & \text{H---C---OPO}_3\text{H}_2 & & \text{H---C---OPO}_3\text{H}_2 \\
| & & | & & | \\
\text{H} & & \text{H} & & \text{H}
\end{array}
$$

Dihydroxyacetone phosphate   *Phosphoglyceric aldehyde*   *1,3-Diphosphoglyceric aldehyde*
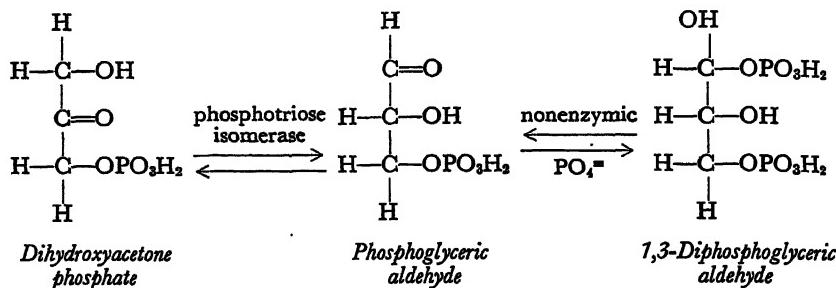
Whether or not the reaction proceeds in exactly this manner is immaterial, as the subsequent reaction has been definitely established and theoretically is of much importance in the caries problem.

(6) Formation of 1,3-Diphosphoglyceric Acid. So far, in the degradation of glucose to alcohol and/or lactic acid very little energy change has taken place. In the next step, a high-energy phosphate bond is produced, of sufficiently high energy content to rejuvenate the ATP that has expended energy for the reaction.

$$
\begin{array}{c}
\text{oxidase} \\
\textit{1,3-Diphosphoglyceric} + \textit{Diphosphopyridine} \xrightleftharpoons{\text{protein}} \textit{1,3-Diphosphoglyceric} + \textit{Diphosphopyridine} \\
\textit{aldehyde} \qquad\qquad \textit{nucleotide} \qquad\qquad \textit{acid} \qquad\qquad \textit{nucleotide} \\
\text{(reduced)}
\end{array}
$$
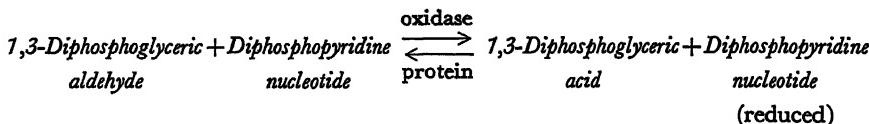
In the above reaction the phosphorylated aldehyde is oxidized by means of diphosphopyridine nucleotide to form the corresponding acid, thus producing a high-energy phosphate bond. Niacin, a member of the vitamin B complex, is the prosthetic group responsible for this reaction. In order for lactic acid to form by means of bacterial fermentation, it is essential that niacin be present; otherwise the series of reactions would cease at the aldehyde stage. The function of niacin in a number of oxidation reactions of the body has been established. The system of oxidation is delicately balanced. Coenzyme I, or the diphosphopyridine nucleotide, is rejuvenated by means of a system involving riboflavin (alloxazine mono-nucleotide and alloxazine adenine dinucleotide). The riboflavin-containing compounds, in turn, pass on the hydrogen to various cytochrome compounds, and finally to atmospheric oxygen to form water. It is interesting to note, at this time, that individuals suffering with pellagra, caused by a lack of niacin, do not ordinarily have dental caries.

(7) The Formation of 3-Phosphoglyceric Acid. The next step in the reaction is nonenzymic. The high-energy compound formed by the oxidation of the aldehyde by means of the niacin-bearing compound is extremely unstable. Thus, when it comes in contact with a phosphate

acceptor, the acceptor is immediately phosphorylated. In this manner the high-energy phosphate bond is rejuvenated.

$$
\begin{array}{ccc}
\overset{\displaystyle O}{\overset{\|}{C}}-OPO_3H_2 & & \overset{\displaystyle O}{\overset{\|}{C}}-OH \\
H-C-OH & & H-C-OH \\
H-C-OPO_3H_2 \;+\; \underset{\displaystyle \longleftarrow}{\overset{\displaystyle ADP}{\longrightarrow}} & H-C-OPO_3H_2 & +\; ATP \\
H & & H
\end{array}
$$

$$
\begin{array}{cc}
\textit{1,3-Diphosphoglyceric} & \textit{3-Phosphoglyceric} \\
\textit{acid} & \textit{acid}
\end{array}
$$
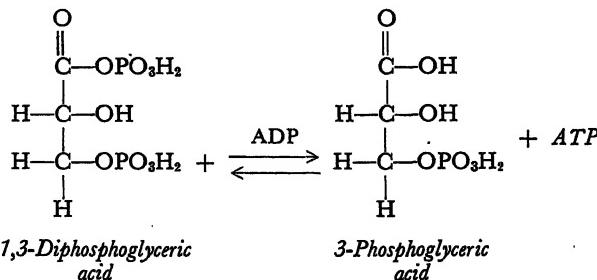
As a result of this reaction the adenyltriphosphate is free to cause the phosphorylation of more glucose under the influence of hexokinase or the synthesis of fructose-1,6-diphosphate from fructose-6-phosphate.

(8) Formation of 2-Phosphoglyceric Acid. After the high-energy diphosphoglyceric acid has been hydrolyzed, the phosphate in the 3 position migrates to the number 2 carbon.

$$
\begin{array}{ccc}
\overset{\displaystyle O}{\overset{\||}{C}}-OH & & \overset{\displaystyle O}{\overset{\|}{C}}-OH \\
H-C-OH & \underset{\displaystyle Mg^{++}}{\overset{\displaystyle \text{triose mutase}}{\rightleftarrows}} & H-C-OPO_3H_2 \\
H-C-OPO_3H_2 & & H-C-OH \\
H & & H
\end{array}
$$

$$
\begin{array}{cc}
\textit{3-Phosphoglyceric acid} & \textit{2-Phosphoglyceric acid}
\end{array}
$$
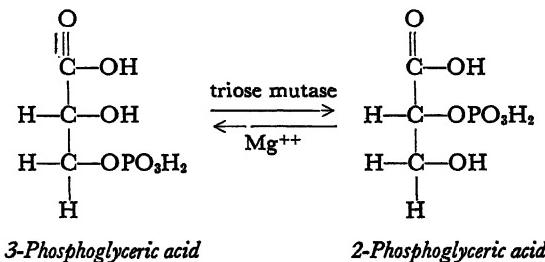
This reaction is carried out under the influence of triose mutase, a protein that is formed in muscle tissue and has been isolated from yeast. It has been established that this enzyme requires a metal ion, Mg^{++}, and perhaps some other metal ion for its activity.

(9) Formation of Phosphopyruvic Acid. The next step in the reaction involves a loss of water, with the formation of an unsaturated acid ester.

$$
\begin{array}{ccc}
\overset{\displaystyle O}{\overset{\||}{C}}-OH & & \overset{\displaystyle O}{\overset{\|}{C}}-OH \\
H-C-OPO_3H_2 & \underset{\displaystyle \longleftarrow}{\overset{\displaystyle \text{enolase}}{\longrightarrow}} & C-OPO_3H_2 \;\;+\; H_2O \\
H-C-OH & & C-H \\
H & & H
\end{array}
$$

$$
\begin{array}{cc}
\textit{2-Phosphoglyceric acid} & \textit{Phosphopyruvic acid}
\end{array}
$$
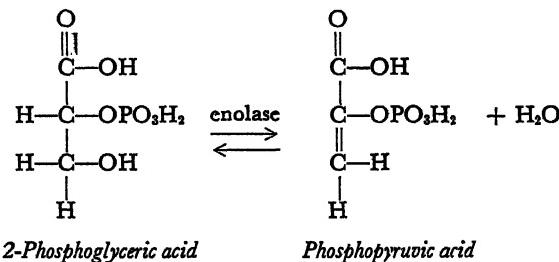
The enzyme system has been isolated in crystal form from yeast and muscle tissue. Magnesium has been found to be an integral part of the enzyme molecule. Among all of the reactions concerned with the formation of lactic acid, this system is the most sensitive to the fluoride ion. Fluorides in very small concentration will inactivate the enzyme. It is thought that the mechanism of the fluoride inhibition is the formation of a magnesium-fluoro-phosphate complex with the enzyme protein.

(10) Formation of Pyruvic Acid. At this stage in the series of reactions for the formation of lactic acid, the phosphate radical leaves the system. Under the influence of the enzyme, phosphopyruvate phosphatase, and a protein-ADP-Mg complex, the phosphate radical is transferred to ADP to form enol pyruvic acid and ATP. Thus, another molecule of ADP can be converted to ATP by means of the high-energy phosphate bond formed as a result of the dehydration of 2-phosphoglyceric acid. And this ATP will serve to phosphorylate more glucose or fructose-6-phosphate.

$$
\begin{array}{cccc}
\overset{\displaystyle O}{\underset{\|}{C}}\text{—OH} & & \overset{\displaystyle O}{\underset{\|}{C}}\text{—OH} & \overset{\displaystyle O}{\underset{\|}{C}}\text{—OH} \\
\text{C—OPO}_3\text{H}_2 & \xrightarrow[\substack{\text{phosphopyruvate}\\\text{phosphatase}\\\text{(Mg-protein)}}]{\text{ADP}} & \text{C—OH} & \text{C=O} \quad + \quad ATP \\
\text{C—H} & + & \text{C—H} & \text{H—C—H} \\
\text{H} & & \text{H} & \text{H}
\end{array}
$$

Phosphopyruvic acid   *Enol pyruvic acid*   *Ketopyruvic acid*
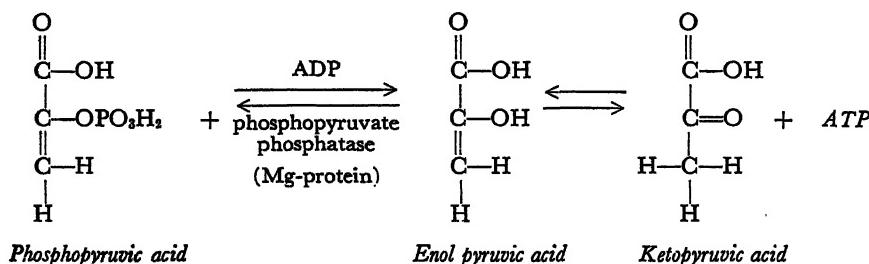
Probably the key compound, in the degradation of sugar in all biological systems is pyruvic acid. However, at this point carbohydrate metabolism in bacteria differs from that in higher animals. Disregarding exceptions for some bacterial and animal forms, one might say that, in general, bacteria convert the pyruvic acid to other similar acids such as lactic, butyric, propionic acids, or to alcohols, such as ethyl, propyl or butyl alcohols; in animals, the general reactions for subsequent products of carbohydrate metabolism involve the aerobic conversion of pyruvic acid to carbon dioxide and water. The latter reactions involve the so-called tricarboxylic acid cycle and require oxygen, ultimately derived from hemoglobin, for completion. However, lactic acid does form in animal tissue, particularly if the muscular activity is great, and in considerable quantities because the anaerobic phase of carbohydrate metabolism is more rapid than the aerobic phase. Insofar as dental caries is influenced by the aerobic phase of carbohydrate metabolism, this stage will not be considered. Suffice it to say that by far the greater part of the energy of metabolism is derived from this phase, and thus

animal metabolism is, in general, far more efficient than microbial metabolism.

(11) The Formation of Lactic Acid, Alcohol and Other Products of Bacterial Metabolism. The acidogenic bacteria and yeasts that have been so far studied contain all the enzymes necessary to convert glucose to pyruvic acid. Many of the organisms also have a full complement of coenzymes and accessory materials, while others lack such fundamental substances as members of the vitamin B complex, and hence with these bacteria the missing factors must be supplied in the media.

After the reactions have proceeded to pyruvic acid, the different enzyme systems of different microorganisms become effective. Thus, yeast will ordinarily convert pyruvic acid to alcohol, while *B.acidophilus* will convert it to lactic acid. Similarly, the propionic acid bacteria [76] and the butyric acid bacteria [92] will convert the pyruvic acid to propionic and butyric acids, respectively. A few of these reactions have been adequately investigated, but most of the systems have been sadly neglected.

The formation of ethyl alcohol under the influence of yeast enzymes is as follows:

$$
\begin{array}{ccccc}
\begin{matrix} O \\ \| \\ C\text{—OH} \\ | \\ C{=}O \\ | \\ H\text{—C—H} \\ H \end{matrix}
&
\xrightleftharpoons[\text{decarboxylase}]{\text{carboxylase}}
&
\begin{matrix} H\text{—C}{=}O \\ | \\ H\text{—C—H} \\ | \\ H \end{matrix}
&
\xrightleftharpoons[\substack{\text{coenzyme I}\\ \text{or}\\ \text{diphosphopyridine}\\ \text{nucleotide}\\ \text{(reduced)}}]{\text{reductase}}
&
\begin{matrix} H \\ | \\ H\text{—C—OH} \\ | \\ H\text{—C—H} \\ | \\ H \end{matrix} + PDN
\\
\textit{Pyruvic acid} & & \begin{matrix}\textit{Acetaldehyde}\\ +\\ \uparrow\\ CO_2\ \text{(gas)}\end{matrix} & & \textit{Ethyl alcohol}
\end{array}
$$
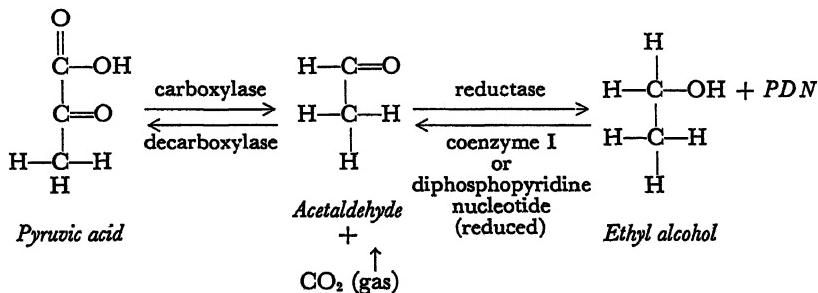
It should be noted that the above reactions are reversible; however, under the normal conditions this is not observed. The carbon dioxide is a gas and hence escapes from the solution, causing the reaction to proceed to the right. In this connection it is of interest to note that another member of the vitamin B complex is active in promoting this metabolic process. Cocarboxylase has been identified as thiamin phosphate, the diphosphoric ester of vitamin B_1. This enzyme is also necessary in the initial phases of the aerobic phase of carbohydrate metabolism in muscle tissue, and hence if the vitamin is absent, an accumulation pyruvic acid and lactic acid is present in the muscle and nerves of individuals suffering from vitamin B_1 deficiency.

The yeast organism possesses the proper enzymes for the production of lactic acid. Under ordinary conditions this system is either not operative or the carboxylase and cocarboxylase are sufficiently potent to

decarboxylate the pyruvic acid before much lactic acid is produced. In any event, if the carboxylase—cocarboxylase system is made inoperative, or decarboxylation is prevented by chemical means, large quantities of lactic acid can be prepared purely by yeast enzymes.

In muscle tissue, and under the influence of *Lactobacillus acidophilus*, very little alcohol or acetaldehyde is formed. This reaction has been studied quite extensively and is as follows:

$$
\begin{array}{cc}
\underset{\text{Pyruvic acid}}{
\begin{array}{c}
\text{O} \\
\parallel \\
\text{C—OH} \\
\mid \\
\text{C}=\text{O} \\
\mid \\
\text{H—C—H} \\
\mid \\
\text{H}
\end{array}}
&
\underset{\text{Lactic acid}}{
\begin{array}{c}
\text{O} \\
\parallel \\
\text{C—OH} \\
\mid \\
\text{H—C—OH} \\
\mid \\
\text{H—C—H} \\
\mid \\
\text{H}
\end{array}}
\end{array}
$$

diphosphopyridine nucleotide (reduced) ⇄ protein oxidase   + *Diphosphopyridine*
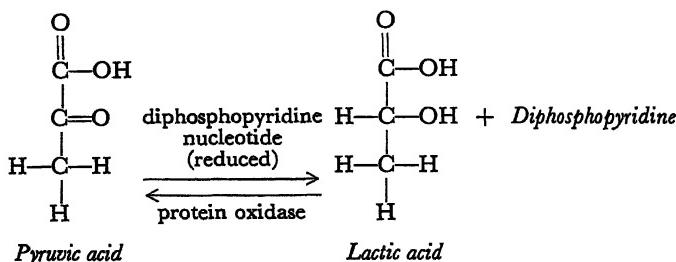
The coenzyme necessary for this reaction is the same as that required to promote the reduction of acetaldehyde under the influence of the yeast enzymes. The oxidase, however, is specific.

Other reactions have been partially studied; the evidence indicates that propionic acid is produced by the reduction of lactic acid. Butyric acid is formed by an aldol condensation of acetaldehyde followed by an intramolecular oxidation-reduction.

Acid Production in the Mouth. The above series of reactions directing normal carbohydrate metabolism in muscle, nerve tissue, microorganisms, and even in germinating seeds is ordinarily called the anaerobic phase of carbohydrate metabolism. Under optimum conditions the reactions are extremely rapid. In fact, these reactions were until recently described as the "explosive conversion of glycogen to lactic acid" which occurs during muscle contraction. In normal muscle tissue the production of lactic acid is extremely rapid.

Before it was recognized that the above scheme of reactions was operative in the production of acids in the mouth, there was an abundance of evidence that should have suggested this. In 1937 it was shown that an important difference between saliva from caries-active and caries-immune individuals is the speed with which sugars are converted into acids.[33]

In 1936 the theoretical aspects of the situation were reviewed[93] but it was not until 1938[94] that actual proof of the matter was at hand. At this time, by the use of available experimental methods, the intermediate compounds formed in the series of reactions were isolated from uninoculated freshly collected mixed saliva when it was allowed to act on sugar. When Stephan[37] first demonstrated the speed with which

TABLE 6. VARIATION IN ACID PRODUCTION BY FERMENTING SUGAR-SALIVA MIXTURES

Reaction Allowed to Proceed to Completion

Patient	Calcium in Saliva (mg/100 cc) Before	After	pH Before	After
Immune:				
1	11.5	103.7	6.4	4.6
2	8.7	47.8	6.6	5.4
3	8.6	61.5	6.8	5.2
4	8.5	53.6	6.8	5.2
5	8.5	42.7	7.0	5.2
6	10.1	35.0	7.2	5.4
Susceptible:				
1	8.9	96.7	7.0	4.6
2	8.2	46.5	6.8	5.2
3	6.5	70.5	6.4	4.4
4	9.3	25.3	6.8	5.2
5	6.1	38.7	6.8	5.4
6	7.9	52.5	7.0	5.4
7	9.0	48.8	7.0	5.4
8	6.7	65.0	6.4	4.6
9	10.0	88.8	7.2	4.8
10	9.3	49.5	6.6	5.0

Reaction Restricted to Four Hours

Patient	Calcium in Saliva (mg/100 cc) Before	After	pH Before	After
Immune:				
1	8.8	8.8	7.4	7.4
2	7.7	4.5	7.4	7.4
3	7.6	4.5	7.4	7.2
4	8.4	12.5	6.8	6.6
5	6.8	6.8	7.0	6.8
6	5.8	7.7	7.2	7.2
Susceptible:				
1	7.8	18.5	7.0	6.6
2	7.7	19.6	6.8	6.6
3	6.5	14.6	6.2	6.0
4	7.4	16.5	7.2	6.6
5	8.7	14.9	6.8	6.6
6	6.6	24.6	6.8	6.6
7	6.4	19.2	6.8	7.0
8	6.2	20.5	7.4	6.6
9	6.2	14.5	7.4	6.8
10	6.8	17.5	7.4	6.8

acids were formed, there was a sound theoretical basis for the fact. It was then realized that the reaction in the mouth for the production of acid from fermentable sugars was in reality slow in comparison to what it might be under the influence of an optimum enzyme system. We are indeed fortunate that the enzyme system in the mouth, even in cases of rampant caries, is in reality a sub-optimal system. One of the prime

variables in caries activity is the efficiency of the enzyme system. In caries-active mouths the system is more optimum for acid production and in immune mouths it is extremely poor. It should be emphasized, however, that even in the most immune mouths the system is present and is capable of producing acids, even though the acids cannot be produced rapidly (Table 6).[33] From the information available, it is clear that the carious process consists of short periods of acid formation accompanied by short intervals of decalcification. Each time a fermentable sugar is allowed to remain in the mouth, acids are formed, and hence a separate "attack" of dental caries may occur at each ingestion of sugar.

Specific Organisms of Dental Caries

At the turn of the century, primarily as the result of Pasteur's influence, specific organisms were being found for many specified diseases, and since dental caries was considered at least partially a bacterial disease, it was only natural that a search would be made for the type of bacteria which would cause dental caries. In 1900 Goadby [95] isolated the Gram-positive bacillus from carious dentine, and he considered that this organism, along with *Streptococcus brevis* may be responsible for the carious lesion. In 1915 Gies [96,97] introduced a series of papers discussing the Miller theory of dental caries, showing that scrapings from the teeth quite frequently contain various forms of streptococci, and non-sporing bacilli capable of producing large amounts of acids were isolated. These were described as aciduric bacilli, the same as *B. acidophilus*, which had previously been described by Moro.[98] Fifty-eight strains of *B. acidophilus* were isolated, one of which was a very potent acid-producer. In addition to the *Bacillus acidophilus*, Gies and his co-workers studied many other forms of bacteria which were frequently found in carious lesions. In 1922 Rodriguez [99] described the *Lactobacillus odonoliticus*, which he had isolated from deep layers of carious dentine, and found them capable of producing pH values ranging from 2.9 to 3.9. He indicated the belief that the *Lactobacillus odonoliticus* is responsible for dental caries. McIntosh and collaborators [100-102] isolated similar bacteria and found essentially the same results as Rodriguez. They recognized that the *Bacillus odonoliticus* may be of several types. In 1925 Bunting [103,104] started a series of investigations in which he found that the *Bacillus acidophilus*, which was probably the same as the *Bacillus acidophilus odonoliticus* of the former investigator, was responsible for dental caries. By the use of a highly specific culture media he found that the *Bacillus acidophilus* was the most predominant organism associated with dental caries. It was his early opinion that this organism was re-

sponsible for dental caries,[105,106] but in the light of recent work it has become obvious that other organisms may be concerned with the carious process. Nevertheless, the work of Bunting, Jay, Crowley and Hadley [107,108] has shown without a doubt that *Bacillus acidophilus* is the most predominant organism associated with dental caries when cultured on a modified Culps tomato-agar medium. There have been other investigations which seem to indicate that other types of aciduric organisms may show the same relationship if a selective media could be found for its growth.

Since the monumental work of the Michigan group, there have been scores of workers who have more or less corroborated these findings and have shown that although the *Bacillus acidophilus* count is not a 100 per cent accurate criterion of caries activity, the correlation between caries activity and the acidophilus count is as accurate as any biological test would be expected to be. Enright, Friesell, and Trescher [14] examined fifty subjects for lactobacilli and found that of fifteen caries-active cases, nine were consistently positive with respect to *Bacillus acidophilus*. Johnston, Williams, *et al.*,[109] studied the bacterial flora of thirty-nine children and found that when the caries activity was very progressive, *Bacillus acidophilus* organisms were continuously present, and of twenty-seven individuals in which caries was not progressing, these microorganisms were continuously absent in eighteen cases, and only sporadically present in eight, and in only one case of the twenty-seven immunes was the acidophilus continuously present. Rosebury and Waugh [110] studied sixty-nine caries-free Eskimos and found that 85 per cent of these individuals had no *Lactobacillus acidophilus* in the saliva. In examining another group of Eskimos, they found that positive cultures were obtained from 80 per cent of the caries-active cases, while they were present in only 13 per cent of those individuals who were caries-free. Snyder [111] also found that in only eleven of 159 caries-free children were *Lactobacillus acidophilus* organisms present.

Becks, Wainwright, and Young [112] found that 77 per cent of caries-free individuals had zero counts of acidophilus, while three of sixty-two caries-free individuals had acidophilus counts as high as 10,000 per cc. Of sixty-six individuals who had active caries, only four held zero counts and two had very low counts; the rest had counts ranging up to 300,000 bacteria. Collins, Jensen, and Becks,[113] using 366 University of California students, found that of ninety-nine caries-free individuals, 81 per cent had zero acidophilus counts, while in rampant caries, 88 per cent had high acidophilus counts and only 12 per cent had zero acidophilus counts.

In view of the large mass of evidence, it is now generally recognized that the acidophilus count, whether or not it has any relation to the

mechanism of dental caries, is an excellent bacteriological test for caries activity, and in all probability those non-medicative procedures that produce a lowering of the acidophilus count will also cause a reduction in caries activity.

There is considerable evidence that other bacteria may play an important role in dental caries. In 1915 Niedergesäss [144] found that acid-forming cocci were always present in deep caries and that they were extremely high acid formers. He regarded these organisms as one of the main causes of dental caries. Hartzell and Henrici,[115] after a comprehensive review of the literature, and on the basis of their own experimental work, said that streptococci are the microbic agents causing dental caries. These organisms are consistently found in the deeper layers of the carious lesion.

Tucker [116] plated scrapings of the surfaces of the teeth of 422 children into an acid broth, pH 5, and found that streptococci of various types were isolated most frequently. He also found that *Lactobacillus* and *Streptococcus albus* developed in this acid media. He concluded, however, that the acid streptococci were found more frequently in immune cases than *Lactobacillus acidophilus*. Anderson [117] found lactobacilli in twelve of sixty-five cases with varying degrees of caries activity, but found acid streptococci in fifty-five of these cases. He concluded that the acid streptococci were more frequently found in caries-active cases than lactobacilli.

Bibby and Hine [118] examined smears made from carious lesions of forty-four individuals and found considerable variation in the bacterial flora. They came to the conclusion that the relation of lactobacillus to carious activity was not as significant as suggested by Bunting. In 1946 Permar, Kitchen and Robinson [119] made a very comprehensive study of the fluctuations of the acidophilus counts in the different samples of saliva from the same individuals. They concluded that although trends in relation to dental caries were apparent, any conclusions as to daily fluctuations were unjustified.

Thus, we can see that many organisms have been correlated with the carious process, and even though *Lactobacillus acidophilus* may not be the predominant organism in a carious lesion, the consensus of opinion is that the salivary counts of *Bacillus acidophilus* are rather accurate criteria of caries activity. In all probability, all of the organisms play a very important role in the carious process. In view of the complicated mechanism of acid-formation in the mouth and the many bacteria that are involved, it is no wonder that so much confusion exists concerning a specific organism for dental caries.

Enzyme Systems in Metabolism. When one considers the problems of metabolism in general, whereby carbohydrates, fats, and proteins are

converted to energy and the various end products, and when one realizes the tremendous number of enzyme systems and accessory activators and conditions necessary, the similarities in living organisms are surprising. It is realized that most forms of life require most of the above constituents for the life process, but they preferentially utilize carbohydrates for the production of energy. Thus man, the highest animals, and most common bacteria prefer to utilize sugars and other carbohydrates for the purpose of supplying energy. Fortunately, in man and animals carbohydrates can be metabolized so that most of the energy can be extracted. On the other hand, although most microorganisms prefer carbohydrates for their energy requirements, they cannot burn them entirely to carbon dioxide and water, and hence the enzyme systems in the microorganisms are fewer and simpler than in man.

Just as man needs many food accessories which he cannot manufacture within himself, so bacteria and lower forms of life require certain food accessories. Thus in the metabolism of carbohydrates by either man or bacteria, the enzymes previously mentioned during the discussion on the mechanism of acid formation must be present for the production of acids from carbohydrates. To present a simple example, man requires niacin, riboflavin and a number of other vitamins for the enzyme systems responsible for normal metabolic processes. Bacteria also require these same food accessories. Man cannot synthesize a sufficient quantity of these, so he must include them in his diet. Many bacteria have the ability to synthesize niacin, riboflavin, thiamin, and other necessary substances. Although the yeast organism needs niacin, riboflavin and thiamin for its normal metabolic processes, it can synthesize these substances and hence it is not necessary to furnish them in yeast media. Thus, the chemical individuals in the human "vitamin B complex" are not vitamins to the yeast organisms. Many organisms, particularly the *Lactobacillus acidophilus* of various strains, cannot synthesize the vitamin B complex, but must be supplied with these substances, and biological assays for members of the vitamin B complex are based upon this fact. Therefore, we would expect the various breeds and strains of bacteria to differ in their food and environmental requirements, just as animals vary in their food and environmental requirements. Just as vitamin C is a vitamin to humans and not to rats, vitamin B is a vitamin to *B. acidophilus* and not to yeast. In man the rate of metabolism is dependent upon a very delicately balanced system of substrate, enzyme systems, vitamins and hormones; so in microorganisms, the rate of metabolism is dependent upon these factors, though probably on a less elaborate scale.

Considering these complexities of living cells, it is no wonder that we have not been able to point out the specific bacteria responsible for

dental caries, if we deal with morphological characteristics of the bacteria. However, if we are interested in one particular phase of metabolism, carbohydrate metabolism to lactic acid, it is obvious that the bacteria having the optimum enzyme system available for the production of acid will be the ones primarily responsible for dental caries. Insofar as the speed of acid production is the prime factor involved, the bacteria having the optimum enzyme system and capable of growing under mouth conditions would be responsible for caries.

The amount and speed of acid formation by means of various bacteria have been determined many times. In general, the rate of acid production by any strain of acidogenic bacteria depends on the conditions under which they are grown. By a careful selection of a culture medium that will furnish all of the food essentials for the particular bacterium (provide all of the enzyme systems in which it is deficient), the maximum speed of acid production by that particular strain of bacteria can be measured. Data from these experiments yield little useful information, as the saliva and other conditions in the mouth may not be ideal. Furthermore, we never find pure-strain organisms in the mouth. Many kinds and varieties live in symbiotic relationship in all mouths. Furthermore, it is quite possible that since several separate reactions are involved in acid production, some organisms may catalyze certain of these reactions rapidly, without having optimum enzyme systems for the others. Since all reactions are necessary, the rate of acid production would then depend entirely on the speed of the slowest reaction.

It was therefore thought that if the enzyme system of each of the common mouth organisms were studied, the offending organism or combination of organisms could be determined. To date the systems of yeast (mouth),[120] *Lactobacillus acidophilus*,[121] *Sarcina lutea*,[122] *Aerobactor aerogenes*,[123] *Staphylococcus albus*,[124] and *Candida albicans*[125] have been studied. Other organisms have been partially studied, but as yet the information is incomplete.

For this work the organisms were isolated from caries-active human mouths in pure strain. After the proper identification of the organisms, they were grown in large numbers, and were either weighed and used in the moist condition, or were dried and used on a dry basis. In some cases the enzymes extracted from a known weight of organisms were used without the cells. The usual procedure was to take each step in the series of reactions for carbohydrate metabolism and permit a known weight of organisms or enzyme concentrate to act upon the substrate. The reaction was prevented from continuing past the step under examination. At the proper time intervals the products of reaction were

determined, and hence the comparative speed of reaction of each step was noted for each organism.

(1) Organisms in the Phosphorylation of Glucose. When glucose is phosphorylated, a mixture of glucose-6-phosphate, fructose-6-phosphate, and fructose-1,6-diphosphate is formed. If the amount of phosphate utilized is known and the amount of mono- and diphosphates are determined, the actual amounts of each sugar phosphate can be calculated. However, for the purpose at hand only the amounts of sugar phosphorylated will be given. As can be seen from Table 7, the organism that causes the phosphorylation of glucose the most rapidly is yeast, while *B. acidophilus* and *Staphylococcus albus* were the least efficient in this respect. At the pH normally found in the mouth yeast, *Sarcina lutea*, *Aerobactor aerogenes* and *Candida albicans* can cause more reaction gram for gram in four hours, than *Staphylococcus albus* and *Lactobacillus acidophilus* can in twenty-four hours. This is extremely

TABLE 7. FORMATION OF HEXOSE PHOSPHATES

(% Formation)

	Time (hrs.)	at pH 5.4	at pH 6.7	at pH 8.6
Staphylococcus albus	24	15	11	12
Lactobacillus acidophilus	24	2	17	30
Yeast	4	35	90	87
Sarcina lutea	4	0	25	0
Aerobacter aerogenes	4	14	22	15
Candida albicans	4	45	50	58

interesting, as it is known that *Lactobacillus acidophilus* is an extremely slow acid former and hence it may well be that the reason for this is that the first reaction is very slow. It is also interesting to note that at pH 6.7, or the pH likely to be in the mouth, most of the organisms produce at a maximum.

In all cases, with the exception of *Lactobacillus acidophilus*, most of the phosphorylated sugar was found to be fructose-1,6-diphosphate. This would indicate that the hexokinase system and the hexoseisomerase were very active in all cases except for this organism. In the case of *Lactobacillus acidophilus*, both reactions were extremely slow, but of that part which did not react there was considerable glucose-1-phosphate. This indicates that the organism is deficient in hexokinase.

(2) Organisms in the Formation of Phosphoglyceric Acid. The next step in the reaction is the dismutation of the hexose diphosphate to the triose phosphates and subsequent oxidation to the phosphoglyceric acid. For the purpose of this paper the data on the formation of the triose phosphates will be omitted, as in the early work these compounds were not isolated, so that the quantitative data on all of the above organisms

is not available. In all cases the phosphoglyceric acid was isolated. The data are given in Table 8. Here again it is observed that *Lactobacillus*

TABLE 8. FORMATION OF PHOSPHOGLYCERIC ACID

	at pH 5.4	at pH 6.7	at pH 8.6
Staphylococcus albus	2	4	4
Lactobacillus acidophilus	3	4	3
Yeast	5	15	15
Sarcina lutea	5	42	1
Aerobacter aerogenes	8	15	12
Candida albicans	3.0	4.0	1.5

acidophilus is rather deficient. Only three of the organisms tested have a potent enzyme system with these reactions. At the pH of the mouth *Sarcina lutea* was by far the most active, while yeast and *Aerobactor aerogenes* were quite as active. Here again the normal pH of the mouth was the most conducive to these reactions. It is interesting to note that with most organisms pH 8.6 was the next most favorable pH for the reactions.

(3) Organisms in the Formation of Pyruvic Acid. The next step in the series of reactions is the dehydration and dephosphorylation of the phosphoglyceric acid. It is quite difficult to stop the reaction quantitatively at this point, particularly if lactic acid is not found. Furthermore, in some cases, particularly in yeast, acetaldehyde is likely to form. The

TABLE 9. FORMATION OF PYRUVIC ACID

	at pH 5.4	at pH 6.7	at pH 8.6
Staphylococcus albus	*	*	*
Lactobacillus acidophilus	5	18	30
Yeast	40	40	60
Sarcina lutea	5	43	9
Aerobacter aerogenes	2	3	2
Candida albicans	17	17	17

* Under the conditions of the experiment, the pyruvic acid was immediately converted to other compounds .

results are shown in Table 9. Yeast, *Sarcina lutea* and *Lactobacillus acidophilus* were the most active in this reaction at the normal pH of the mouth. It should be noted that this reaction proceeds best at pH 8.6, but that much activity is present at pH 6.7. Here we find that *Lactobacillus acidophilus* may carry out the reaction with considerable speed, so it is quite possible that if the other organisms would form the phosphoglyceric acid for the *Lactobacillus acidophilus,* it could proceed to form lactic acid very rapidly.

A very interesting observation is that in no case was pyruvic acid isolated from reaction mixtures in which *Staphylococcus albus* or *Staphyloccus albus* enzyme concentrate was used. It was found that

as soon as the pyruvic acid was formed it polymerized to form a diketoaldehyde and a diketo acid of considerable interest in fat metabolism. It would be theoretically possible to use *Staphylococcus albus* in the prevention of caries if it were in sufficient concentration to destroy the pyruvic acid as fast is it was formed or before the enzymes from the other organisms could convert it to lactic acid. In any event, the resulting compounds would be a mixture resulting from the competitive action of all the pyruvic-acid-activating enzymes.

(4) Organisms in the Formation of Lactic Acid. At this point in the sequence of events it is possible for the pyruvic acid to be decarboxylated to form acetaldehyde, and subsequently alcohol. It can be polymerized or it can be oxidized or reduced. In muscle tissue it will condense with oxaloacetic acid in preparation for complete oxidation to carbon dioxide and water. It is interesting to note that in the case of *Staphylococcus albus,* yeast, and *Candida albicans,* at the pH that normally exists in the mouth, no lactic acid is formed. The results are given in Table 10.

TABLE 10. FORMATION OF LACTIC ACID

	at pH 5.4	at pH 6.7	at pH 8.6
Staphylococcus albus	0	0	0
Lactobacillus acidophilus	3	7	4.7
Yeast	0–1 *	0–6.5 *	0–5.9 *
Sarcina lutea	1	1	1
Aerobacter aerogenes	0.2	0.1	0.2
Candida albicans	0	0	0

* Under conditions not conducive to decarboxylation.

As can be seen from the table, *Lactobacillus acidophilus* is the most active organism in the formation of lactic acid. In all of the other cases, with the exception of the mouth yeast, little or no lactic acid is formed.

If the reaction mixture is so adjusted that recarboxylation is difficult, yeast readily forms lactic acid. This would be expected, as the system for reducing acetaldehyde to alcohol is quite similar to that in which pyruvic acid is reduced to lactic acid.

Combined Enzyme Systems in the Mouth. Considering all of the reactions and microorganisms studied, it is clear that no one organism by itself would be a rapid acid-former. However if the enzymes elaborated by one organism could be utilized by another growing in symbiotic relationship, the combination of the proper organisms should result in a rapid production of acid. On the basis of the above, yeast and *Lactobacillus acidophilus* should be such a combination. The yeast could produce the phosphorylated compounds, and if the carboxylase-cocarboxylase system in the yeast was not sufficient to use up the

pyruvic acid before the enzymes of the *Lactobacillus acidophilus* could, then lactic acid could be produced at a much faster rate. This system has been studied along with a few others with theoretical results. It is probably not a coincidence that yeast and *Lactobacillus acidophilus* are found so frequently in caries-active saliva. It is hoped that when all of the mouth organisms have been studied from this point of view we can predict which are the most effective in producing dental caries.

The Mechanism of the Carious Process

On the basis of the above evidence, a theory was proposed [126] which offered a rational explanation for most of the anomalies that had previously existed. It explained why successful methods of caries control were successful, and furnished suggestions concerning new methods of control which were subsequently found successful.[127,128] This theory has in general been accepted as the true mechanism of dental caries,[1] and with few modifications in the light of recent research is as follows.

Acids are formed in all mouths whether the individual is susceptible or immune to caries. There is a wide difference in the rate of acid formation in different mouths and in the same mouth or areas of the same mouth. Even with this wide variation, all teeth would eventually decalcify if the acids were not destroyed by the oral secretions. Thus, whether or not decalcification occurs depends upon whether or not the acids formed can be neutralized, dissipated, or otherwise destroyed before sufficient time has elapsed for decalcification to start or continue. All mouths have a natural tendency to neutralize, dissipate, or otherwise destroy acids, but the tendency varies from mouth to mouth, and even in different areas of the same mouth. As a matter of fact, the tooth itself, the last barrier, is a powerful neutralizer, and hence any acid on the surface of a tooth is self limiting.

Classification on the Basis of Caries Activity. Thus, it is obvious that the two main variables controlling caries activity are the rate of acid formation and the rate of acid neutralization. There are a multitude of factors that may either directly or indirectly regulate the two prime variables, but in the final analysis these two variables control the carious process, and on the basis of these factors all cases may be classified, within reasonable limits, into one of four groups:

(1) Individuals with rapid oral acid formation and rapid oral acid neutralization.

On the basis of examination and analysis of the saliva, it would seem that most of the individuals in this country belong to this class. Individuals who have an average occlusion, no oral anatomical defects, an average supply of saliva, a fair diet such as is prevalent in this country,

and a not excessive sugar intake, will belong to this class. Under these conditions perhaps one to three cavities will appear each year. The caries rate will, of course, vary somewhat, as the most inacessible areas of the mouth, such as pits and fissures, will decay before a period of semi-immunity to caries will result.

(2) *Individuals with rapid oral acid formation and slow oral acid neutralization.*

It is obvious that individuals of this type are those suffering from rampant caries. Only a small percentage of the population of this country falls into this category, but this group furnishes a large pro- portion of the difficult dental cases. Individuals in this group are usu- ally characterized by a relatively high sugar intake, poor occlusion, poor contacts, poor dental anatomical relations, usually a poor diet, or an acid-ash diet, and very frequently a subnormal supply of stimu- lated saliva. These cases are rather difficult to control because of the very rigorous limitations on the individual's habits. However, by a careful analysis of the variables, the causative factors can be found, and if the patient is cooperative, control of caries-activity can be achieved.

(3) *Slow oral acid formation and rapid oral acid neutralization.*

In this country about two to five per cent of the population belong to this class. These individuals are usually characterized by perfect oral anatomical relations, a rather copious supply of stimulated saliva, a well-balanced alkaline-ash diet, and in general a self-cleansing mouth. These individuals in general do not care for sugar, but occasionally those with an exceptional self-cleansing mouth may consume consider- able fermentable sugar without causing a high caries activity. Occasion- ally an individual of this type is found who practices no oral hygiene procedures, never has professional dental care, and never bothers about his diet, but still is caries-free. This type of individual very frequently will develop pyorrhea.

(4) *Slow oral acid formation and slow oral acid neutralization.*

So far as the author is aware, few, if any, of the individuals in this country belong to this class. From theoretical considerations the indi- vidual should be characterized by a diet containing no sugar or fer- mentable carbohydrate. Furthermore, the diet should normally be defi- cient in vitamins, calories, and many of the food accessories. Or if the individual is extremely deficient in niacin, the sugar intake would be immaterial. These individuals should be in about the same category as the class (1) individuals in respect to the caries activity, but would probably be somewhat less caries-active. They would have little or no protection against acid decalcification, but since no acid is formed, no protection is needed. It is the opinion of the author that the indi- viduals in war-torn Europe, such as were described by Schour,[129] would

belong to this class. The people suffering from pellagra, as described by Mann, and the malnourished, starving individuals in India should belong to this class. It would be interesting to examine these individuals and analyze their saliva.

With the above in mind it is obvious that if one knew all of the variables that would regulate the rate of oral acid formation and the rate of oral acid neutralization, one would have a clear picture of the carious process. Naturally, there are many well-known factors that would alter, regulate or modify these two prime variables; however, they are very numerous and they differ from mouth to mouth so that it is some-
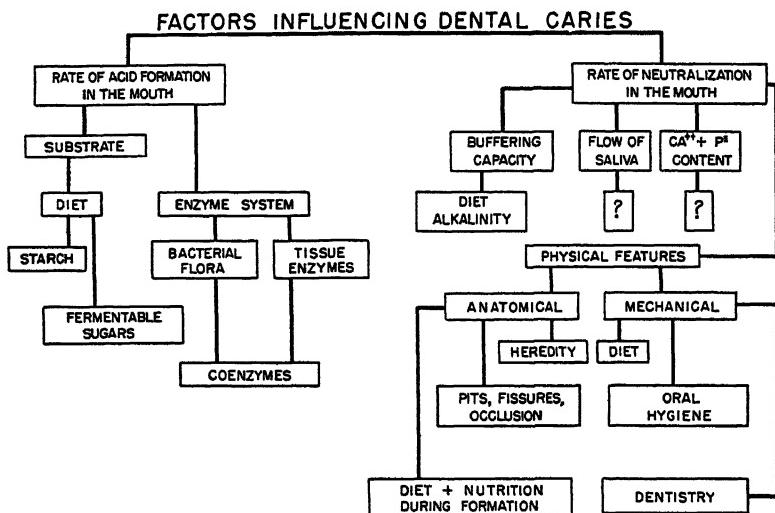
Figure 3. Factors influencing dental caries.

times difficult to determine the exact factors involved in the individual case. Even if the factors are determined, one cannot always assign exact quantitative values to each variable. However, many variables have been determined which are general to a greater or lesser degree in all mouths, and which are readily ascertained. We do know the variables that regulate acid formation. This is for the most part a purely chemical phenomenon and can be determined with accuracy. The factors that regulate acid neutralization, on the other hand, are not so well understood and cannot, in many cases, be ascertained with accuracy. We can, however, discuss the general factors and point out, from theoretical considerations, the line of investigation for any individual. These factors are tabulated in Figure 3.

Factors Regulating the Rate of Acid Formation. The factors that regulate the rate of acid formation are relatively well known. The reac-

tions taking place during bacterial fermentation have been described previously in this paper. On the basis of these studies it is clear that the two and only two governing factors are (1) an adequate substrate and (2) the proper enzyme system. Under the conditions that exist in the mouth there are many conditions that would control each of these two factors. It must be emphasized that in order to have any acid, both an adequate substrate and enzyme system must be present.

The Factors that Control the Substrate. First of all, the substrate must be in the mouth or in the site of action for a sufficient period of time for the reactions to proceed to acids. Furthermore, it must be in the form in which it may be converted to acids. The substrate is taken into the mouth, e.g., during the mastication of foods. Thus it ordinarily comes entirely from the diet. Furthermore, it consists only of that portion of the diet that remains in the mouth after the food is masticated and swallowed. This is indeed a very important fact. The diet as such exerts a controlling influence in that it helps to regulate the neutralizing factors in the mouth by means of the saliva, but it should be kept in mind that all of the acid which serves as the decalcifying agent is derived *only* from the excess food that remains in the mouth after mastication and swallowing. This factor is probably the main reason why, under starvation conditions, with little or no protection against caries, few caries are produced. It is quite possible that if food were taken in gelatin capsules or by stomach tube, no caries would result, regardless of the type of diet. This treatment is not recommended as it would undoubtedly result in much periodontal disturbance, but it would be an interesting experiment.

Another important factor is the type of substrate. From theoretical considerations and under the conditions that exist in the mouth, the polysaccharides, starches, dextrins, glycogen; the disaccharides, maltose, sucrose, lactose; and the monosaccharides, glucose, fructose, mannose, and galactose, should be capable of forming acids. The fats and proteins should be of no consequence. From actual clinical observations, only the disaccharides and monosaccharides are of much importance. In 100 per cent of the individuals in groups (1) and (2), caries can result from these sugars. In only about 10 per cent of the cases, mostly those in group (2), can caries be derived from starches. When one considers the theoretical aspects of the situation, one finds that the clinical observations are substantiated. As previously described, the conversion of glucose to lactic acid requires some fifteen steps, most of which are governed by an enzyme and coenzyme system. For the sake of brevity, let us designate these reactions by letters, rather than the actual compounds. Thus, if we designate glucose as A, glucose-6-phosphate as B, etc., then the equation will be a series, as A→B→C→D→E, etc. In all

cases of the two classes of patients in which acids are formed rapidly, the whole sequence of reactions may require only a few minutes at the most. One must remember that the speed of the production of the final product, lactic acid, depends solely upon the slowest reaction of the sequence. Thus, if the reaction A→B required five seconds, and all the rest required five seconds, with the exception of the reaction C→D, which we will say requires fifteen minutes, then the fastest time that we would expect the end product to appear is approximately fifteen minutes. Under these conditions the over-all reaction would be slow. It is generally agreed that the conversion of starches to acids requires first a hydrolysis of the starch to maltose, with a subsequent further hydrolysis to glucose, as shown in the following reaction.

$$Starch \xrightarrow{\text{ptyalin}} Maltose \xrightarrow[\text{maltase}]{} Glucose$$

Most of the oral bacteria can convert maltose to glucose at a very rapid rate, and hence maltose will produce acids under the conditions that exist in a "rapid acid formation mouth" almost as fast as glucose. However, the reaction

$$Starch \xrightarrow{\text{ptyalin}} Maltose$$
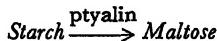
in most mouths is quite slow. In most cases saliva must remain in contact with starch for several minutes before any reducing sugars are detectable. Thus, under ordinary conditions the reaction

$$Starch \longrightarrow Maltose \longrightarrow Glucose \longrightarrow B \longrightarrow C \longrightarrow etc. \longrightarrow Lactic\ acid$$
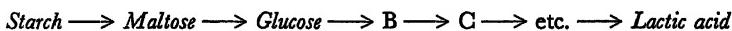
is quite slow. Unless the mouth in question has an extremely low protection factor, then the acids from starches are of no consequence. However, in those cases where a high ptyalin activity is present, and at the same time a low protective factor is present, starches may furnish the acids for the decalcification. The clinical evidence supports this view, and it may be that the observations of Schour in Italy are due to this factor. In Italy during the war the main foodstuff was spaghetti, and other starch foods, but dental caries were extremely infrequent. Laboratory evidence also supports this view.[130]

Another factor that controls the substrate is the anatomical features of the mouth. Although sugars and starches are taken into the mouth in the diet, the amount of these substances retained in the mouth is not necessarily a function of the amount consumed. If the mouth has anatomical defects, such as pits and fissures, poor occlusion, poor contacts, and other situations that would prevent a self-cleansing mouth, much of the ingested food will be retained. However, if the mouth is self-cleansing, then little of the food will be retained. Thus, some indi-

viduals may consume relatively large quantities of sugar and still be relatively "caries-immune" while others may eat small quantities of sugar and still retain much of it and have "rampant caries."

The retention of food and sugar in the mouth is not wholly a function of the self-cleansing ability of the mouth. The character of the food is of considerable importance. Some foods cling to the teeth and readily become impacted, while the coarse fibrous foods are less likely to be retained in the mouth. Furthermore, the amount of saliva would be a factor in washing the soluble sugars from the mouth.

Another factor in the retention of substrate in the mouth is the frequency of eating. If sugar is ingested at each meal, there is the possibility of three separate "attacks" of dental caries. Some individuals will eat many times per day, and the between-meal snacks are very frequently sweets in some form. This constant eating may well explain, in part, the increase in caries activity during the active growing period in children and young adults. Furthermore, confusion exists today concerning the relation of caries activity to pregnancy. It has been observed that when obstetricians limit the food intake during pregnancy, no change in caries activity results.

From a theoretical point of view, oral hygiene could play a very important role in regulating the substrate in the mouth. It is obvious that if the maximum acid potential is about twenty minutes after the ingestion of food, oral hygiene procedures could, if applied before this time, remove much of the substrate and even some of the acids if they are already formed. This procedure is being tried experimentally, and the preliminary results are very favorable. Even a discreet rinsing of the mouth with water while still at the table should be of considerable benefit.

On the basis of the above, it is quite evident that there are a large number of variables that regulate the substrate in the mouth. There are undoubtedly others that would apply to specific cases.

The Factors that Regulate the Enzyme System. It is known that all of the enzymes necessary for the production of lactic acid and other organic acids are elaborated by the microbial flora of the mouth. From a study of the enzyme systems of six oral microorganisms, it is obvious that each bacterium has a qualitatively different system, some having a higher or lower concentration of the different enzymes. As pointed out above, yeast has a high concentration of the enzymes required for the first few steps of the degradation process, but a low concentration of the enzymes required for the conversion of pyruvic acid to lactic acid; *Lactobacillus acidophilus* has a low concentration of enzymes necessary to start the series of reactions, but a very high concentration of the enzymes necessary to convert pyruvic acid to lactic acid. Thus,

when both organisms grow in symbiotic relationship there is the prob-
ability that the combination can form acids very rapidly. It is quite
possible that by the proper combination of organisms a truly efficient
enzyme system can be present. The exact combination awaits further
research.

The problem is not as simple as it would seem. It is possible that a
combination of organisms which could furnish an optimum concen-
tration of each of the necessary enzymes could be found, and under the
conditions in the mouth still not form acids at a maximum rate. There
is considerable evidence to indicate that the reactions may be intra-
cellular as well as extracellular. Thus, it would be possible to have
some of the enzymes bound into the cell in such a manner that they
would not operate with extracellular substrates. This possibility is un-
likely, as all of the reacting substances are generally diffusible through
most cell membranes.

All the enzymes do not necessarily originate in the oral flora. It is
quite possible that some of the enzymes come from the saliva or oral
tissues. The ptyalin so necessary for the conversion of starch to fer-
mentable sugars is derived from the saliva. It has been demonstrated
that the dentine can furnish a phosphatase which may be effective
in phosphorylating sugars. However, in general, few of the necessary
enzymes have been demonstrated in the saliva, and until they have
been found there, it is unnecessary to consider them as an important
factor.

Some of the coenzymes and activators have been demonstrated in the
saliva. Furthermore, some of the bacteria have been demonstrated as
being deficient in some of the coenzymes. *Lactobacillus acidophilus* is
known to be deficient in the vitamin B complex, but yeast is extremely
high in these factors, Thus, as long as the yeast is present the system
would not lack coenzyme I. There is considerable evidence that mem-
bers of the B complex are secreted in the saliva. If this is the case,
some of the vitamins would depend upon the diet. The fact that
people suffering from pellagra do not ordinarily have caries would
tend to substantiate this contention. The saliva also furnishes the
metallic ions, primarily magnesium, that are so necessary as activators
in many of the enzyme systems. These, in turn, would also be a func-
tion of the diet.

In view of the above, it is quite clear that the enzyme system can
and does come from the microbial flora. The efficiency of the system
would be dependent in part on the kind and number of bacteria pres-
ent. The enzyme system including the coenzymes and activators may
come from the saliva and would in part depend upon the diet of the
individual. The diet of course, particularly that portion which remains

in the mouth after the ingestion of food, also affects the bacterial flora. It is not known how much of this effect is due to the coenzymes and activators or how much is due to the quantity and type of food that remains in the mouth.

Another factor that may assume important proportions in some cases is that of enzyme inhibitors. All of the enzymes are proteins and hence any substance that will denature or change the molecular configuration of proteins should inactivate the enzymes. Under ordinary conditions nothing of this nature would be found in the mouth. In addition to the above effect, certain substances are specific inhibitors for specific enzymes. An example of this is the fluoride ion. The fluoride ion will selectively inhibit enolase and block the formation of pyruvic acid. Whether this is the main mechanism of the action of fluorides on dental caries is not known, but it can be assumed that this action plays some role in the process. Other possibilities along these same lines are those of bacterial antagonism and the competition of enzymes. In the first case it is possible that some organism will form substances that will interfere with the enzymes of others. With our recent knowledge concerning the mechanism of antibiotics, this possibility is interesting.

In regard to the "competition" of enzymes for substrate, other very interesting possibilities are suggested. For instance, it is known that *Staphylococcus albus* converts pyruvic acid to non-acid products, and that *Lactobacillus acidophilus* converts pyruvic acid to lactic acid. Thus, if the enzyme system of *Staphylococcus albus* were in large excess, little if any lactic acid would be formed. Whether or not this mechanism is ever operative is not known. Since tissue enzymes will convert pyruvic acid to carbon dioxide and water, it is possible that if the proper enzymes were present the pyruvic acid could be converted to harmless substances. This system may partially explain the observed antagonism of caries activity and pyorrhea.

In summary it might be stated that the enzyme system probably depends primarily on the bacterial flora, which in turn is dependent to a large extent on the diet. Some of the necessary requirements must be present in the diet and some originate in the saliva. In addition to this there is the possibility that in some cases other factors, such as enzyme inhibitors, bacterial antagonism, and a competition of enzymes for substrates, may alter the path of the reactions.

Factors Influencing the Neutralization or Dissipation of Acids. Insofar as acids are formed from carbohydrates retained in the mouth, it is obvious that the factors that influence the dissipation or neutralization of these acids are of prime importance, and in many cases the determining factor of dental caries. There are many conditions found in the average mouth that would tend to protect against the action of acids.

Perfectly formed and anatomically correct teeth and mouth and a copious supply of highly buffered saliva would afford the maximum protection. Which of the factors that regulate neutralization are the most important quite possibly depend upon the individual.

The Flow of Saliva. From theoretical considerations one of the most important variables would be the amount of saliva that can be formed under the stimulation of the usual mastication of foods. The basal flow of saliva should be of little importance, as the damage is done during the period in which the saliva would normally be stimulated. Stimulation of saliva during the mastication of foods is a natural protection against acid formation. The amount of saliva so produced varies tremendously from individual to individual, and according to the results of Trimble,[131] it is almost a straight-line function of caries activity. In cases where no saliva is produced, rampant caries is always present. One of the main observations in the absence of saliva is the exceptional amount of dental caries present. In most mouths, however, the differences in caries activity between a copious flow of saliva and a moderate flow of saliva is not so great. This is presumably due to the protective action of other influences of the mouth. The difference between a moderate flow of saliva and no saliva is very great. The amount of saliva present affects its distribution. The protective action of saliva is due to several factors, e.g., washing or diluting action, and buffering action. If the saliva is present during the formation of acid, then the buffering capacity is important. The amount of saliva secreted basally and under stimulation is a function of the sympathetic and parasympathetic nervous system. One of the main characteristics of salivary flow is that a stimulation of the sympathetic system usually causes a decreased output of saliva. Thus, a highly nervous individual or an individual under an emotional strain should produce a lesser quantity of stimulated saliva. This may explain the clinical observation that during times of emotional stress dental caries become more prevalent. Many attempts have been made to cause an increase in the flow of saliva, either by drugs or by physical means. In no case has success along these lines been attained.

The Buffering Capacity of Saliva. All saliva is normally highly buffered, though many individual differences may exist. In 1915 Marshall [132] of California discovered that individuals immune to dental caries had saliva with a higher buffering capacity than the individuals suffering from rampant caries. The buffering capacity of the saliva is governed primarily by metabolic processes and the diet. An alkaline-ash diet will usually produce a more highly buffered saliva, while a diet low in alkaline-ash will usually produce saliva with a low buffering capacity.[133] The buffering capacity, even when extremely low,

cannot be increased more than 100 per cent by the ingestion of an alkaline-ash diet. Furthermore, as previously stated, the buffering capacity of the saliva would be of no consequence if for any reason it was prevented from reaching the areas in which acid is forming. Many factors may cause the saliva to be deflected from the areas in which it is needed. The calcium-phosphorus content of the saliva could be considered as a buffering effect. If the calcium-phosphorus content of the saliva is completely ionized into calcium ion and tertiary phosphate, the saliva is always supersaturated and the ionic calcium and phosphorus must be depressed further or utilized by the acid before the teeth are attacked. Here again we have little information concerning the regulation of the calcium and phosphate ion in saliva. The total calcium content of the saliva is usually characteristic of the individual, and may vary spontaneously in the same individual. The calcium and phosphorus content of the saliva, however, cannot be altered by the ingestion of calcium, phosphorus and vitamin D, unless the individual is so deficient in these elements that it is reflected in low calcium and phosphorus in the blood stream. This condition is extremely unusual in this country.

The Physical Characteristics of the Mouth. Probably the most important factor, aside from the amount of saliva, is the physical characteristics of the mouth. The physical characteristics of the mouth may be either anatomical or mechanical. By the anatomical aspects we mean the position of the teeth, the shape of the arches, the shape of the teeth, and the shape of the jaws and the occlusion and contacts of the teeth. It may categorically be said that any anatomical anomalies that would impede the free flow of saliva to all surfaces of the teeth would prevent the saliva from performing its normal function of neutralizing acids. In addition to this, any anatomical feature that would aid in the accumulation of food stuffs and their retention during the act of mastication would provide more suitable substrate for acid formation. The very fact that one consumes sugars and carbohydrates does not necessarily mean that these materials are retained in the mouth. Thus, the anatomy of the mouth is a governing factor as to whether acids are actually formed and whether they are neutralized or otherwise dissipated.

The factors that govern the anatomical aspects of the mouth are predominantly hereditary. There is some evidence to indicate that diet and nutrition during formative periods of the teeth and during the growth of the individual is important, but it is generally conceded that in this country heredity is of greater importance, and may explain most of the trends associated with dental caries.

In addition to the anatomical aspects of the problem there are many purely mechanical impediments such as fillings and orthodontic appliances which may, if improperly placed, furnish areas that may not be

readily accessible to the saliva. Obviously these may hinder the normal neutralizing influences of the mouth.

The Dental Plaque. One of the most important and controversial factors that would tend to prevent the contact of saliva with the surfaces of the teeth and the acids of decalcification is the so-called "dental plaque" or "mucin plaque." This mechanical barrier is present on all teeth, whether susceptible or immune to caries. This deposit or film that exists primarily in the "susceptible" areas of the teeth has been investigated and discussed since the conception of the chemicoparasitic theory of dental caries. Despite all the attention it has received, there is still no agreement concerning the composition, the source, actual appearance and function of this material. In all probability, part of the confusion concerning the dental plaque is due to differences in the plaque itself, or to different conceptions concerning its nature.

Several investigators [134, 135, 136] have proposed that the plaque itself is capable of decomposing to form acids in sufficient strength to decalcify the teeth. Whether or not acids may be derived directly from the plaque has not been satisfactorily answered, but it has been shown that the plaque material is an excellent nutrient for bacterial growth.[137] Kirk [138] was of the opinion that the plaque is the major agency for the localization of all carious lesions. This opinion does not conform to that of Miller [139] and Pickerill,[140] who did not believe that the plaque is essential to dental caries. In general, however, it is thought that the mucin plaque is of major importance and most studies [141-143] seem to indicate that at least in caries-susceptible individuals this is quite probable. There is also considerable evidence to support the idea that some plaques are not harmful. It was suggested by Bibby [144] that the dental plaque may not be a uniform material, and that some types may be beneficial while others may be harmful. This latter concept is more in accord with the findings of Bradel and Blayney.[143]

Despite all of the excellent work on this subject there is no agreement concerning the source, composition, and function of this material. Kirk [138] was able to form plaques from salivary mucin and was of the opinion that the plaque on the teeth was derived from salivary mucin. Most investigators, however, feel that the major portion of plaque material consists of living and dead microorganisms.[141-143,145-147]

Although there has been a tremendous amount of bacteriological investigation of the plaque, relatively little chemical work has been done. Lothrop,[148] on the assumption that plaques are primarily mucin, made a rather intensive study of salivary mucin, a glycoprotein with a variable composition. The nitrogen content varied from 9 to 14 per cent. Inouye [149] found salivary mucin to contain 10.45 per cent nitrogen. An actual analysis of plaques on human teeth showed the material to be a

protein containing from 7.75 to 11.2 per cent nitrogen. This is of about the same order as analysis of salivary mucin.

Campaigne [150] investigated mucin plaques on human teeth, synthetic plaques, and salivary mucin. He described two types of plaques: one rather gray, tough and rubbery; the other white and very soft. An analysis of these plaques indicated the same composition but a different hydrogen-ion concentration. The white, soft plaque, commonly called "materia alba" had a reaction comparable to that of saliva. The gray rubbery ones were more acid. It is quite possible that these two types contain the same material, different only in regard to the proximity to the isoelectric point.

It has been found that when the teeth are thoroughly cleansed, all the plaque material can be removed from the surface of the teeth. If there is no ingestion of food or other disturbances in the mouth, the plaque in the form of a thin, readily removed film will rapidly regenerate. Obviously, this material is not derived from the food, and hence it must be derived from the saliva. In all probability it is derived from bacterial action on salivary proteins. The analysis would not indicate whether the plaque is derived from the mucin or from the bacterial masses, for they are both protein and contain comparable amounts of nitrogen.

Regardless of the source and composition of the dental plaque, it would seem that anything so consistently present in the biological system should have some function. Furthermore, it would seem that anything so generally present would have a useful function, providing it had not been disturbed, altered or changed, by external factors.

The material is extremely difficult to study from a chemical point of view, primarily due to lack of sufficient material for analysis. In 1939 Campaigne [151] developed a glass electrode with which it is possible to measure the H^+ of as small a quantity as 0.001 ml of material. About this same time a micro method [152] for the determination of lactate was formulated. With these two tools it was possible to determine the buffering capacity of plaque and carious material. If the pH and pK_1 are known, then the ratio of neutralized acid to free acid can be calculated. If the total acid is known, then the amount of neutralized acid and free acid can further be calculated by means of the equation

$$pH = pK_1 + \log \frac{BA}{HA}$$

The above methods showed that the plaque material from caries-immune individuals had the highest buffering capacity of any substance in the body. Plaques from caries-active cases had an excellent buffering capacity, but far less than that of the caries-immune plaques.

It is quite possible that the buffering capacity of plaque material is due to a multiplicity of actions, e.g., to the buffering action of its constituent protein, that of hydrolyzed amino acids, or perhaps some adsorbed base or salts. The first possible substance that may be suggested is ammonia or ammonium salts. Kesel has considerable information that ammonia may play an important role in the prevention of dental caries. The ammonium concentration in the saliva of most caries-immune individuals is not sufficient to explain the immunity. Kesel has suggested that perhaps the ammonia has been absorbed in the plaque. This is a very plausible suggestion, as one would expect a nitrogenous substance like certain proteins to absorb large quantities of ammonia, or ammonium salts. Another possibility would be that the plaque contains protein or protein fragments or amino acids that are predominantly basic. The amino acid content of the plaque material has never been determined. This would be an extremely interesting and important project. Yet another possibility is that free amino acids are present that would act as powerful buffers, but there is no information concerning them. This explanation is unlikely because of the insolubility of the plaque material.

Whether or not an adequate explanation can be made, it is obvious that the buffering capacity of plaque material is very considerable, and may be a very important factor in the prevention of dental caries. Furthermore, it explains why acid foods do not ordinarily cause decalcification. It is quite possible that the plaque is a natural defense against the decalcifying action of such substances as all fruit juices. It also suggests the possibility that dental "erosion" in which a typical non-carious decalcification occurs, may be due to a faulty formation of non-carious plaques.

The Formation of the Carious Lesion. Practically all of the research concerning the actual progress of the decalcification and formation of the cavity has been histological in nature. Several attempts have been made to duplicate the natural process in vitro but to date the attempts have not been very successful. When acid is confined on the surface of the tooth in vitro by mechanical means, decalcification does occur as would be expected; however, in only rare occasions is it possible to obtain many of the true characteristics of a natural carious lesion. In most instances the cavity produced by acid decalcification is merely a "cupping out" of the inorganic phase without the finger-like projections along the rods to the dentine, and only in very rare instances is the inverted-cone type of progress observed. The cavity produced by acid decalcification in vitro strongly resembles the decalcification ordinarily observed in cases of "opalescent dentine" or those cavities that have occurred in non-vital or pulpless teeth, after the tooth has been

"treated." From a physicochemical point of view, this is exactly what one would expect.

It has been shown many times that the inorganic phase of the tooth is definitely an ionic compound and that the crystal lattice is orientated into a definite pattern. It has further been shown that the tooth is permeable to various ions and in some cases to relatively large molecules. Thus, there is definitely a transfer of ions under certain conditions. Radioactive salts are known to penetrate the enamel, the fluoride ion will definitely replace the OH ion,* and there is a tendency for water to transfer as exhibited by various dye experiments.

In view of the above, the vital tooth should be considered as a semi-permeable membrane, and there is a tendency for ions and water to pass from the blood stream of the pulp to the saliva or vice versa. This "membrane," the tooth, however, is not an ordinary membrane. It is ordinarily very dense, it is practically impermeable to most substances, and it is, so far as the enamel is concerned, almost entirely ionic in nature, so that the laws of diffusion through a membrane would be materially affected. Under these conditions the only substances that would be able to pass the membrane unmolested would be the un-ionized molecules such as water and certain small molecular weight, un-ionized compounds, such as ether, alcohol and perhaps a few dyes. If the inorganic portion of the enamel were un-ionized, then other substances such as the fluoride ion, sodium ion, and hydrogen ion could pass at will, but as this is not the case, those ions that will replace the calcium, phosphate, or hydroxyl ion will be stopped by effecting such replacement in the enamel. Many of these ions will alter the chemical properties of the enamel. The fluoride ion will make the enamel more insoluble. If the hydrogen should replace the calcium, then a water-soluble compound is formed. In the true carious process this process takes place.

The Role of Osmotic Pressure. Whether or not ions or molecules will diffuse through a membrane depends largely upon the concentrations of the solutes on both sides of the membrane, which in turn is reflected in the osmotic pressure of the respective fluids. In the normal mouth the osmotic pressures approximately balance: the osmotic pressure of the blood inside the tooth and pulp is about 6.8 atmospheres. The osmotic pressure of the saliva is also about 6.8 atmospheres. Under these conditions the transfer of water and other molecules is at a minimum. However, if very soluble materials are taken into the mouth, such as sugars or certain salts, this transfer would be very great. A simple

---

* The relation between hydroxyapatite and fluorapatite is shown by the x-ray work of C. A. Beevers and D. B. McIntyre [*Mineralog. Mag.*, 27, 254 (1948)]. Fossil bone gradually accumulates fluorine.

calculation $(PV=nRT)$ would demonstrate this. If an 18 per cent glucose solution were taken into the mouth, the osmotic pressure of the solution is 22.4 atmospheres. Thus, there is a tendency for the sugar to diffuse toward the pulp and there is a tendency for the pulp to become dehydrated by the outward passage of water. The phenomenon could be compared to the action of red blood cells when placed in a hypertonic solution.

In the normal sound tooth this action is not evident because most teeth are so impermeable that the actual passage of molecules would not take place even with these tremendous osmotic-pressure differentials. In many cases such as when candy is masticated, sugar concentrations of much greater concentrations may result, and no manifestations of these differentials would be exhibited. However, if a tooth is made more permeable because of the carious process or because of defects in the tooth such as pits, fissures or lamellae, the dehydration of the pulp may be evident as demonstrated by a "sweet tooth." Most foodstuffs would not penetrate the intact enamel and even sugar would have no harmful effect if nothing further occurred, but in the case of sugar, further action does occur under ordinary mouth conditions.

The Role of the Hydrogen Ion. It has been conclusively shown that when fermentable sugars come in contact with the enzyme systems of the mouth they are rapidly converted to acids, primarily lactic acid. The prime characteristics of all acids is that they will in the presence of water form hydrogen ions, H^+. Thus, part of the tremendous osmotic pressure of sugar solutions in the mouth may be due to hydrogen ions. Furthermore, the osmotic pressure will be increased by the fermentation process. Let us consider the 18 per cent glucose solution again. If this sugar were converted quantitatively to lactic acid, then the osmotic pressure on the surface of the teeth would be slightly greater than 44.8 atmospheres (672 lb per square inch), and the difference in osmotic pressure between the tooth surface and the blood stream of the pulp would be tremendous. Furthermore, hydrogen ions would be concentrated on the surface of the tooth and would be literally sucked in toward the pulp. The hydrogen ion is so small that it could, if not stopped by the calcium ions of the crystal lattice, pass right through to the pulp. Because of the ionic structure of the enamel this cannot take place. The hydrogen ion would *start* to penetrate until it was replaced by a *calcium* ion. Thus, the hydrogen ion would not penetrate far because of the calcium ion, but because of the replacement, the portion of the crystal replaced would be water-soluble and hence would wash out, leaving a submicroscopic hole in the tooth. The hydrogen ions would necessarily follow the lines of least resistance. In the case of a lamellae

it is conceivable that the hydrogen ion may penetrate almost entirely through the enamel before it is "clipped off" by a calcium ion.

In view of the above, the chemical process would be as follows:

(1) Acids are formed in the dental plaque or on the surface of the tooth by the fermentation of sugars.

(2) The acids thus formed or the hydrogen ions thus formed are ordinarily neutralized by the buffering action of the plaque or saliva and are hence rendered innocuous. If the plaque or saliva does not neutralize them because of the large amount of acid or the inefficiency of the plaque or saliva, then they are held on the surface of the tooth.

(3) Under the influence of the osmotic-pressure differential, there is a tendency for ions to go toward the pulp and water to diffuse from the pulp. Insofar as the hydrogen ions are the smallest ions known, they would naturally travel along the lines of least resistance toward the pulp. If lamellae were present, this would be the line of least resistance. If lamellae were not present then they would penetrate along the least dense or least calcified rod. It is inconceivable that all the enamel rods are identical, so the penetration would be faster along some rods than others, but at the surface the decalcification would be as widespread as the area of acid. Under these conditions a cone of decalcification would be observed with the apex at the point of greatest penetration or least resistance to hydrogen ions, and a hole or pathway would be gradually made through the enamel. At first the "hole" is submicroscopic, but as the hydrogen ions pour through it, it is enlarged to the point where bacteria may also be "sucked" in, and if sugar molecules are also "sucked in" the bacteria have plenty of nutrient material to feed on. The process is repeated within the tooth; the discoloration that precedes the bacteria simply indicates areas of partial decalcification with the liberation of organic material. As bacteria are drawn into the "hole" the enzymes of the proteolytic type can act upon this decalcified organic material and produce discoloration.

(4) When the hydrogen ions get through the enamel they no longer have to follow the very confining path. The organic material in the dentine will permit diffusion sideways as well as toward the pulp and another cone-shaped area is produced. By this time the path toward the pulp is well defined and decalcification can occur at good speed. This is contrary to the observations on the dentine of extracted teeth. In the vital organ the organic material does not impede the progress of the hydrogen ion, but in the extracted tooth where no osmotic pressure differences exist, the organic material impedes decalcification.

Thus, from a physicochemical point of view, the vital tooth should decalcify more rapidly and with a different pattern than a treated tooth or an extracted tooth. Furthermore, with this concept there is no

fundamental difference between caries starting from a sound tooth sur-
face and caries starting from lamellae. Furthermore it affords for the
first time an excellent theoretical explanation for unilateral caries.
Thus, if one approximating surface has lamellae while another does
not, the condition is inevitable.

The Control of Dental Caries

We have now explained in some detail the mechanism whereby
dental caries occur, including the way acids form and the way they
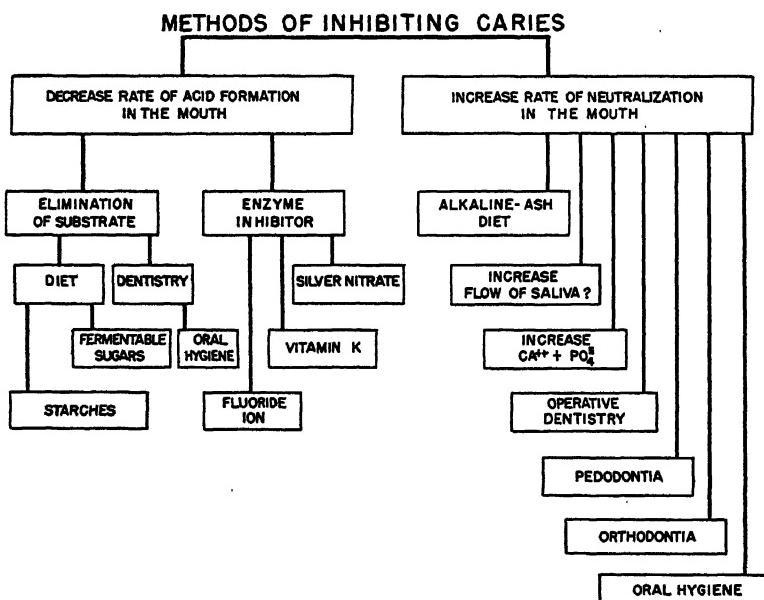
Figure 4. Methods of inhibiting caries.

attack the inorganic portion of the tooth. We have further described
the many variables that control acid formation and the normal dissipa-
tion of these acids. Thus, from a theoretical point of view, it should be
comparatively easy to control dental caries. All that would be necessary
would be to (1) prevent acid formation in the mouth, or (2) to increase
the rate of acid neutralization or destruction. This can be graphically
shown, as in Figure 4.

In order to prevent dental caries there are two fundamental methods
of approach. Of these two methods it has been found more practical
to inhibit the formation of acid than to increase the rate of acid neutral-
ization or destruction; however, it may be said that in many procedures

both factors are influenced. For the purpose of explanation it may be well to describe each method separately and then arrive at a procedure which involves the practice of several methods.

The Control of Acid Formation. Since the factors which regulate acid formation are much better known that the factors that control the neutralization of the acids, and since as these factors are more amenable to control, more success in the control of caries has been attained by interfering with acid formation. This can be attained by either controlling the substrate or by controlling the enzyme system.

The Control of the Substrate. Many individuals have attained control of dental caries by the voluntary or involuntary elimination of fermentable sugars from the diet. In some cases this has resulted in a better balanced diet as in the case of the work of Boyd,[153] and hence this line of approach may result in a better quality saliva. In other cases, such as in the starvation diets in certain parts of India,[154] in the war and postwar Europe, in the prisoner of war camps, in Japan and Germany, the low-sugar diets were very poor and unbalanced, and in many cases very deficient in food essentials. In any event, the mere elimination of fermentable carbohydrates in most cases will control caries. It should be emphasized that the elimination of carbohydrates from the mouth is a prime consideration rather than a balanced diet, providing the sugars are essentially eliminated from the diet. Under these conditions no acid can form. Even if the diet is very deficient no acid can form, and no caries can result. If sugar is present in an unbalanced diet, then protection against acids becomes imperative, and hence a well-balanced diet is always recommended.

The elimination of sugars, and in some cases starches, from the diet is not a practical means of controlling caries in this country. The eating habits and the taste for sugar is so firmly entrenched as to make this almost impossible. However, if the dentist can be partially successful along this line the practice is worthwhile.

Other methods of producing a "sugar-free" mouth are possible. Of course, if one must have sugar, it can be taken in gelatin capsules, but that also is not feasible. A more practical method is to remove the sugars or acids from the mouth as fast as they are formed. This is being tried at present with excellent results. If the teeth are brushed immediately after each ingestion of food, the caries activity is materially decreased. If facilities are not available for brushing, a water rinse is very valuable.

The Control of the Enzyme System. Although theoretically the regulation of the diet is an excellent method of caries control, from a practical standpoint interference with the enzyme system is more effective. It is only necessary to have a suitable enzyme inhibitor present

during the period the acids would normally form. This has been tried experimentally with zephiran, urea, urea quinine, synthetic vitamin K and the fluoride ion. In all cases in which the enzyme inhibitors were present along with the sugar, acid inhibition was effected. In all the above cases with the exception of fluorides in the drinking water, the presence of the inhibitor is dependent upon the cooperation of the individual. Under these conditions mass control of dental caries is impossible.

In the case of the fluoride ion, the inhibitor may be always present. Furthermore, the action is multiple in that the solubility of the enamel toward acids is also effected. The experience with fluorides makes it evident that it is not necessary to stop all acid formation. It is only necessary to inhibit the reactions to the extent that the oral environment can take care of the situation. Thus, an inhibition of ten per cent may cause a sixty to seventy per cent decrease in caries activity. This is self-evident with the use of topical application of sodium fluoride.

From a theoretical point of view the ideal situation would be the addition of an enzyme inhibitor directly to sugar, so that all commercial sugar would contain the inhibitor. Under these conditions a true public health prevention of caries would be attained. At present no such inhibitor is available, but of the several hundred inhibitors tested, glyceric aldehyde [155] seems to fill most of the requirements. At present the only recognized method is the topical application of two per cent sodium fluoride. This has been tried and tested by the U. S. Public Health Department.

The Control of the Neutralizing Influences. To date very little success has been attained by this method. There are so many variables in this category that even if some success were attained, it would be extremely difficult to evaluate. One can improve the neutralizing influences of the saliva by a well-balanced alkaline ash diet, but this usually entails a decrease in sugar consumption. For this reason the success attained cannot be accurately evaluated. One cannot influence the amount of stimulated saliva or the calcium or phosphate concentration in the saliva by any known means.

The dentist and orthodontist can practice good dentistry and thus correct the anatomical features that would increase caries activity. This practice has probably been quite effective in this country, but here again one cannot accurately evaluate the exact effect. Good oral hygiene procedures may be quite effective in removing harmful plaques, but here again the cooperation of the individual is important. It is quite possible that when more is known about the dental plaque, such knowledge can be used in the prevention of caries.

Conclusion

Thus, it can be said that although we know about the causes and mechanism of dental caries, we do not as yet have methods of control that can be practical on a public health scale, but we do have excellent methods that can be utilized by the dentist. There are tests available whereby the dentist can evaluate the treatment, and there are many methods at his disposal for preventing caries.

A logical procedure is suggested as follows:

(1) Caries tests should be secured in order to determine the caries activity.
(2) Dietary habits should be learned for the purpose of determining the sugar and starch intake and the adequacy of the diet.
(3) An accurate mouth examination with radiographs should be secured in order to determine the clinical condition and the anatomical and mechanical aspects of the mouth.
(4) The stimulated flow of saliva should be measured in order to determine the amount of saliva available.
(5) The acid-neutralizing power of the saliva should be ascertained in order to determine the buffering capacity and as a check on the diet. With the above information the dentist should be able to determine the factors causing the condition, and thus prescribe the required methods of control.

If the caries activity is high, the sugar intake high, the buffering capacity of the saliva low, and if the anatomical features of the mouth are adequate, a simple correction of the diet would probably suffice. The patient would be required to reduce the sugar intake and include much fruit and vegetables in the diet. This should correct the condition and the decrease in caries activity could be determined by a subsequent caries-activity test.

If the caries activity is high, the sugar intake low, the amount of buffering capacity of the saliva low, and the anatomical features of the mouth poor, the case will be very difficult. The patient should, of course, be placed on a well-balanced, sugar-free diet. He should be instructed to clean his mouth immediately after the ingestion of food. The teeth should be treated with fluorides and perhaps a liquid dentifrice containing fluorides should be furnished. In this case it may not be possible to prevent all caries unless extreme cooperation is obtained, but the caries activity can be reduced to a minimum.

If the dentist understands all of the multitude of variables and trains himself to recognize them, it is definitely possible with the tests and procedures available today to control dental caries. The degree of

success depends upon his knowledge of the process and procedures of control and upon the degree of cooperation he can secure from his patients.

REFERENCES

1. Michigan Workshop on the Evaluation of Dental Caries Control Technics, *J. Am. Dental Assoc.*, **36**, 3 (1948).
2. Miller, W. D., "Der Mikro-organismen der Mundhoele," 2nd ed., Leipzig: George Thieme, 1892.
3. Black, G. V., *Trans. Illinois Dental Soc.*, 180 (1886).
4. Roseberry, H. H., Hastings, A. B., and Morse, J. K., *J. Biol. Chem.*, **90**, 395 (1931).
5. Klement, R., and Trömel, G., *Z. physiol. Chem.*, **213**, 263 (1932).
6. ———, and ———, *Klin. Wochschr.*, **12**, 292 (1933).
7. Bale, W. F., Le Fevre, M., and Hodge, H., *Naturwissenschaften*, **24**, 636 (1936).
8. Wendt, G. L., and Clark, A. H., *J. Am. Chem. Soc.*, **5**, 882 (1923).
9. Holt, L. E., LaMer, V. K., and Chown, H. B., *J. Biol. Chem.*, **64**, 509 (1925).
10. Fosdick, L. S., and Starke, A. C., *J. Dental Research*, **18**, 417 (1939).
11. Kuyper, A. C., *J. Biol. Chem.*, **159**, 411 (1945).
12. Robinson, R., "Significance of Phosphoric Esters in Metabolism," New York Univ. Press, 1933.
13. Benedict, H. C., and Kanthak, F. F., *J. Dental Research*, **12**, 277 (1932).
14. Enright, J. J., Friesell, H. E., and Trescher, M. O., *J. Dental Research*, **12**, 759 (1932).
15. Forbes, J. C., *J. Dental Research*, **13**, 415 (1933).
16. Becks, H., *Abhandl. Geb. Klin. Zahnheilk.*, **20**, 1 (1929).
17. ———, and Wainwright, W. W., *J. Dental Research*, **14**, 387 (1934).
18. ———, and ———, *J. Dental Research*, **17**, 198 (1938).
19. Wainwright, W. W., *J. Dental Research*, **17**, 235 (1938).
20. Becks, H., *J. Dental Research*, **18**, 431 (1939).
21. ———, and Wainwright, W. W., *J. Dental Research*, **18**, 447 (1939).
22. Wainwright, W. W., *J. Dental Research*, **18**, 441 (1939).
23. Becks, H., and Wainwright, W. W., *J. Dental Research*, **20**, 627 (1941); *ibid.*, 637; with Young, D. H., *ibid.*, 171.
24. Starr, H. E., *J. Biol. Chem.*, **54**, 43 (1922).
25. ———, *J. Biol. Chem.*, **54**, 54 (1922).
26. McIntosh, J. James, W. W., and Lazarus, B. P., *Brit. J. Exptl. Path.*, **6**, 260 (1925).
27. Henderson, M., and Millet, J. A. P., *J. Biol. Chem.*, **75**, 559 (1927).
28. Brawley, R. E., *J. Dental Research*, **15**, 55 (1935).
29. ———, *J. Dental Research*, **15**, 79 (1935).
30. ———, *J. Dental Research*, **15**, 87 (1935).
31. Grossman, L. I., and Breckman, B. M., *J. Dental Research*, **16**, 179 (1937).
32. Unpublished work in laboratory of Northwestern University Dental School, Chicago, Ill.
33. Fosdick, L. S., and Hansen, H. L., *J. Am. Dental Assoc.*, **24**, 1445 (1937).
34. Karshan, M. *J. Dental Research*, **15**, 383 (1936).
35. Etherington, J. W., and Trimble, H. C., *J. Dental Research*, **14**, 218 (1934).
36. Stephan, R. M., *J. Dental Research*, **17**, 251 (1938).
37. ———, *J. Am. Dental Assoc.*, **27**, 718 (1940).
38. Fosdick, L. S., Campaigne, E. E., and Fancher, O. E., *Illinois Dental J.*, **10**, 85 (1941).
39. Fancher, O. E., and Fosdick, L. S., *J. Dental Research*, **20**, 267 (1941).
40. Cagniard, L., *Ann. chim. phys.*, **68**, 206 (1838).
41. Schwann, T., *Ann. Physik*, **41**, 184 (1837).
42. Kützing, F., *J. prakt. Chem.*, **11**, 385 (1837).
43. Berzelius, J., *Jahresber.*, **15**, 237 (1836).

44. Liebig, J., *Liebig's Annal. Chem.*, **30**, 250 (1839).
45. Pasteur, L., *Ann. chim. phys.*, (3) **58**, 323 (1860).
46. Traube, M., p. 119, Berlin, Ferd. Dummler's Verlagsbuch, 1858.
47. Berthelot, M., *Compt. rend.*, **50**, 980 (1860).
48. Buchner, E., *Ber.*, **30**, 117 (1897).
49. Harden, A., *J. Inst. Brewing*, **11**, No. I (1905).
50. Young, W. J., *Biochem. Z.*, **32**, 178 (1911).
51. Neuberg, C., and Karczag, L., *Biochem. Z.*, **36**, 68 (1911).
52. Harden, C., and Young, W., *Biochem. Z.*, **40**, 453 (1912).
53. Neuberg, C., and Kerb, J., *Biochem. Z.*, **64**, 251 (1914).
54. ———, *Biochem. Z.*, **71**, 1 (1915).
55. Embden, G., Griesbach, W., and Schmitz, E., *Z. physiol. Chem.*, **93**, 1 (1915).
56. ———, ———, and Laquer, E., *Z. physiol. Chem.*, **93**, 125 (1915).
57. Euler, H., Ohlsen, H., and Johannson, D., *Biochem. Z.*, **84**, 402 (1917).
58. Neuberg, C., *Biochem. Z.*, **96**, 180 (1919).
59. ———, Farber, E., Levite, A., and Schenk, E., *Biochem. Z.*, **83**, 224 (1917).
60. Meyerhof, O., *Z. physiol. Chem.*, **101**, 165 (1918).
61. Robinson, R., *Biochem. Z.*, **16**, 809 (1922).
62. Meyerhof, O., *Biochem. Z.*, **178**, 395 (1926).
63. ———, and Lohmann, K., *Naturwissenschaften*, **14**, 1277 (1926).
64. ———, and ———, *Biochem. Z.*, **185**, 113 (1927).
65. ———, and Gottschalk, A., *Biochem. Z.*, **154**, 492 (1924).
66. Nielson, R., and Sandberg, E., *Biochem. Z.*, **174**, 196 (1926).
67. Neuberg, C., *Biochem. Z.*, **103**, 320 (1920).
68. Meyerhof, O., *Biochem. Z.*, **178**, 462 (1926).
69. Robison, R., and Morgan, W. T., *Biochem. J.*, **24**, 119 (1930).
70. ———, and King, E. J., *Biochem. J.*, **25**, 323 (1931).
71. ———, and ———, *Chem. Industries*, **48**, 143 (1929).
72. Embden, G., and Zimmerman, M., *Z. physiol. Chem.*, **167**, 114 (1927).
73. Harden, A., and Henley, F. R., *Biochem. J.*, **23**, 230 (1929).
74. Meyerhof, O., *Biochem. Z.*, **292**, 25 (1937).
75. ———, and Kiessling, W., *Ber.*, **68**, 243 (1935).
76. Wood, H. G., Stone, P. W., and Werkman, C. H., *Biochem. J.*, **31**, 349 (1937).
77. Neuberg, C., and Kobel, M., *Biochem. Z.*, **272**, 445 (1934).
78. Meyerhof, O., *Nature*, **132**, 373 (1933).
79. Neuberg, C., *Biochem. Z.*, **272**, 445 (1934).
80. Virtanen, A. I., *Soc. Sci. Fennica Commentationes Phys. Math.*, **20**, 1 (1925).
81. Lutnak-Mann, C., and Mann, T., *Biochem. Z.*, **281**, 140 (1935).
82. Euhler, V., and Adler, E., *Z. physiol. Chem.*, **234**, 122 (1935).
83. Dische, Z., *Naturwissenschaften*, **24**, 462 (1936).
84. Meyerhof, O., *Naturwissenschaften*, **23**, 859 (1935).
85. Colowick, S. P., and Kalckar, H. M., *J. Biol. Chem.*, **148**, 117 (1943).
86. Kalckar, H. M., *J. Biol. Chem.*, **148**, 127 (1943).
87. Ostern, P., Guthke, J. A., and Terszakowec, J. Z., *Z. physiol. Chem.*, **243**, 9 (1936).
88. Zakov, N. E., *Biokhimiya*, **6**, 163 (1941).
89. Herbert, D., Gordon, H., Subrahmanyan, V., and Green, D., *Biochem. J.*, **34**, 1108 (1940).
90. Warburg, O., and Christian, *Biochem. Z.*, **311**, 209 (1942).
91. ———, and ———, *Biochem. Z.*, **314**, 1949 (1943).
92. Werkman, C. H., and Wood, H. G., *Botan. Rev.*, **8**, 1 (1942).
93. Fosdick, L. S., and Hansen, H. L., *J. Am. Dental Assoc.*, **23**, 401 (1936).
94. ———, *J. Am. Dental Assoc.*, **26**, 415 (1939).
95. Goadby, K. W., *J. Brit. Dental Assoc.*, **21**, 65 (1900).
96. Gies, W. J., *J. Allied Dental Soc.*, **10**, 137 (1915).
97. ———, Hull, L. M., and Millikin, J. C., *J. Allied Dental Soc.*, **10**, 459 (1915).
98. Moro, E., Wein, *Klin. Wochschr.*, **13**, 114 (1900).
99. Rodriguez, F. E., *Military Dental J.*, **5**, 199 (1922).

100. McIntosh, J., James, W. W., and Lazarus, B. P., *Brit. J. Exptl. Path.,* 3, 138 (1922).
101. ———, ———, and ———, *Brit. J. Exptl. Path.,* 5, 175 (1924).
102. ———, ———, and ———, *Brit. J. Exptl. Path.,* 6, 260 (1925).
103. Bunting, R. W., and Palmerlee, F., *Proc. Soc. Exptl. Biol. Med.,* 22, 296 (1925).
104. ———, and ———, *J. Am. Dental Assoc.,* 12, 381 (1925).
105. ———, Nickerson, G., Hard, D. G., and Crowley, M., *Dental Cosmos,* 70, 1 (1928).
106. ———, ———, and ———, *J. Am. Dental Assoc.,* 15, 1230 (1928).
107. Jay, P., *J. Am. Dental Assoc.,* 27, 393 (1940) .
108. Hadley, F. P., *J. Dental Research,* 13, 415 (1933).
109. Johnston, M. M., Williams, E. H. M., Anderson, P. C., Drake, T. G. H., Tisdall, F. F., and Kaake, M. J., *J. Am. Dental Assoc.,* 23, 1493 (1937).
110. Rosebury, T., and Waugh, L. M., *Am. J. Diseases Children,* 57, 871 (1939).
111. Snyder, M. L., *J. Dental Research,* 18, 497 (1939).
112. Becks, H., Wainwright, W. W., and Young, D. H., *J. Dental Research,* 20, 171 (1941).
113. Collins, R. O., Jensen, A. L., and Becks, H., *J. Am. Dental Assoc.,* 29, 1169 (1942).
114. Niedergesäss, K., *Arch. Hyg.,* 84, 221 (1915).
115. Hartzell, T. B., and Henrici, A. T., *J. Natl. Dental Assoc.,* 4, 477 (1917).
116. Tucker, W. H., *J. Infectious Diseases,* 51, 444 (1932).
117. Anderson, T. G., *J. Dental Research,* 16, 489 (1937) .
118. Bibby, B. G., and Hine, M. K., *J. Am. Dental Assoc.,* 25, 1934 (1938).
119. Permar, D., Kitchen, P. C., and Robinson, H. B. G., *J. Dental Research,* 25, 475 (1946).
120. Fosdick, L. S., and Wessinger, G. W., *J. Am. Dental Assoc.,* 27, 203 (1940).
121. ———, and Starkes, A. C., *J. Am. Dental Assoc.,* 28, 234 (1941).
122. ———, and Calandra, J. C., *Arch. Biochem.,* 6, 9 (1945).
123. ———, and Dodds, A. F., *Arch. Biochem.,* 6, 1 (1945).
124. ———, and Rapp, G. W., *Arch. Biochem.,* 1, 379 (1943).
125. ———, and Blohm, T., in press.
126. ———, *Can. Dental J.,* 9, 249 (1943).
127. ———, *J. Am. Dental Assoc.,* 29, 2132 (1942).
128. Burrill, D. Y., Calandra, J. C., Tilden, E. B., and Fosdick, L. S., *J. Dental Research,* 24, 273 (1945).
129. Schour, I., and Massler, M., *J. Am. Dental Assoc.,* 35, 1 (1947) .
130. Fosdick, L. S., and Rapp, G. W., *J. Dental Research,* 23, 85 (1944).
131. Trimble, H. C., Etherington, J. W., and Losch, P. R., *J. Dental Research,* 17, 299 (1938).
132. Marshall, J. A., *Pac. Dental Gaz.,* 23, 335 (1915).
133. Forbes, J. C., and Gurley, W. B., *J. Dental Research,* 26, 415 (1939).
134. Gore, J. J., *J. Dental Research,* 17, 69 (1937).
135. Lohmann, A., *Dental Cosmos,* 46, 1086 (1904).
136. Gore, J. T., *Dental Cosmos,* 77, 942 (1935).
137. Lothrop, A. P., and Gies, W. J., *J. Allied Dental Soc.,* 5, 262 (1910).
138. Kirk, E. C., *Dental Cosmos,* 52, 729 (1910).
139. Miller, W. D., *Dental Cosmos,* 44, 425 (1902).
140. Pickerill, H. P., "Prevention of Dental Caries," p 186, Baillière, Tindall and Cox, London, 1912.
141. Hemmens, E., Blayney, J. R., and Harrison, R. W., *J. Dental Research,* 20, 29 (1941).
142. Miller, B. E., Muntz, J. A., and Bradel, S. E., *J. Dental Research,* 19, 473 (1940) .
143. Bradel, S. F., and Blayney, J. R., *J. Am. Dental Assoc.,* 27, 1601 (1940).
144. Bibby, B. G., *J. Am. Dental Assoc.,* 22, 222 (1935).
145. Chase, S. W., *Anat. Record,* 33, (5), 357 (1926).
146. Bibby, B. G., and Van Huysen, G., *J. Am. Dental Assoc.,* 20, 828 (1933).
147. Williams, J. L., Gies, W. J., and Bodecker, C. F., *Dental Cosmos,* 69, 123 (1927).
148. Lothrop, A. P., *J. Allied Dental Soc.,* 7, 410 (1912).
149. Inouye, J., *J. Dental Research,* 10, 7 (1930).

150. Campaigne, E. E., and Fosdick, L. S., *Northwestern Univ. Dental Research and Grad. Bull.*, **38**, 27 (1938).

151. Fosdick, L. S., Campaigne, E. E., and Fancher, O. E., *Illinois Dental J.*, **10**, 85 (1941).

152. Miller, B. F., and Muntz, J. A., *J. Biol. Chem.*, **126**, 413 (1938).

153. Boyd, J. D., *J. Am. Dietet. Assoc.*, **18**, 211 (1942).

154. Dais, C. D. M., *J. Am. Dental Assoc.*, **31**, 52 (1944).

155. Fosdick, L. S., and Calandra, J. C., *J. Dental Research*, **26**, 309 (1947).

Part III

TECHNOLOGICAL APPLICATIONS

SONIC AGGLOMERATION OF CARBON BLACK AEROSOLS

C. A. Stokes

Director of Research and Development, Godfrey L. Cabot, Inc., Boston, Mass.

Introduction

Aerosols are disperse systems of liquid or solid particles suspended in a gas. They may result from either dispersion or condensation processes. The particles of aerosols formed by dispersion processes, as for example by dry grinding, are usually larger than about 5 to 10 microns in diameter. Aerosols formed by dispersion are called *dusts* if formed by dispersing solids, and *sprays* if formed by dispersing liquids. This type of aerosol is fairly easy to collect because of the relatively high settling rates of particles greater than 5 to 10 microns in diameter (about 0.5 cm/sec for a 10-micron diameter spherical particle of density 2 gm/cc settling in air and 0.16 cm/sec for a 5-micron particle). On the other hand, aerosols formed by condensation—called *fumes* if composed of solid particles or *mists* if composed of liquid particles—are less than about 5 to 10 microns in diameter and are extremely difficult to collect unless first agglomerated by one means or another. (The settling rate in air of a 1-micron diameter spherical particle of density 2 gm/cc is 0.006 cm/sec.)

It is not possible, of course, to draw a sharp line of demarcation between aerosols formed by dispersion processes and those formed by condensation processes. Aerosols of either type may contain a wide range of particle or agglomerate sizes. Condensation aerosols composed of solids are also called *smokes,* as for example the carbon smoke formed when a hydrocarbon burns with limited air supply. *Fog* is another term for liquid aerosols formed by condensation. Natural water fog, and artificial water fog formed when solid carbon dioxide is dropped in warm water, are examples. Numerous excellent review articles are available on aerosol behavior in general. [1,2] An entire issue of the *Transactions of the Faraday Society* [3] is devoted to this subject and is perhaps the best review on this subject in the English language.

Carbon-Black Aerosol Collection. The collection of carbon-black smoke, whether in the form of furnace-, channel-, lamp- or thermal black, is perhaps the most important single case in all industry in which a fume or smoke is deliberately made and collected as a product. The total United States production of all types of pigment- and filler-grade carbon black is over 600,000 tons yearly. Nearly every type of fume collection equipment yet devised has been used or seriously proposed for use in collecting carbon black. Channel black, like the black made from vegetable oils by the ancient Chinese, is collected by thermal precipitation on cooled surfaces, in this case the steel channels from which the process gets its name. Lampblack is collected in gravity settling chambers, and in some plants the residual smoke is scrubbed with water in a baffled spray tower [4] to recover additional quantities of black. Thermal black is collected in bag filters when it is made as a primary product and in water-filled traps and baffled spray towers when it is made as a by-product of oil-gas or hydrogen manufacture. Furnace black is agglomerated in Cottrell electrostatic precipitators and the agglomerates are collected in cyclones of conventional design.

It is in the furnace-black industry that the most growth and technological progress are taking place today. Furnace black accounts for nearly half of the total carbon black produced in the United States. Recently new grades of highly reinforcing furnace black have been developed that can replace channel black for both quality and economy. The United States produces about 90 per cent of the world's supply of carbon black. There is little doubt that the furnace process will eventually supplant the channel process for producing all except certain special high-priced pigment carbons. This paper is concerned with the application of a new method of aerosol agglomeration to the furnace-black process.

Description of Carbon-Black Aerosols. As initially formed in the partial combustion flame of the channel-, furnace-, or lampblack processes, or in the thermally cracked gas envelope of the thermal black process, carbon black consists of primary particles * varying in number average diameter from about 0.005 to 0.5 micron. For a given type and grade of black the distribution of sizes about the mean is fairly narrow, i.e., about 90 per cent of the particles of any given grade of black will lie within a tenfold range of diameter. Furnace blacks range in mean diameter from about 0.03 to 0.07 micron. Figure 1 shows an electron-microscope photograph of semi-reinforcing furnace black before agglomeration by the Cottrell apparatus. Figure 2 shows the same black after

---

* The primary particle is defined by the smallest particle that can be resolved by the electron microscope. It appears to be spherical in shape when viewed by the electron microscope. The primary particles are always more or less clustered together, giving rise to such descriptive terms as *structure units* and *chain-like aggregates*.
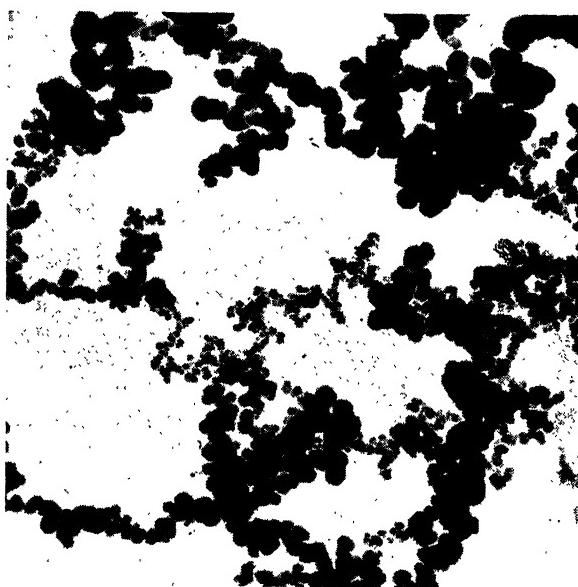
Figure 1. Semi-reinforcing furnace black before Cottrell agglomeration. (25,000x)
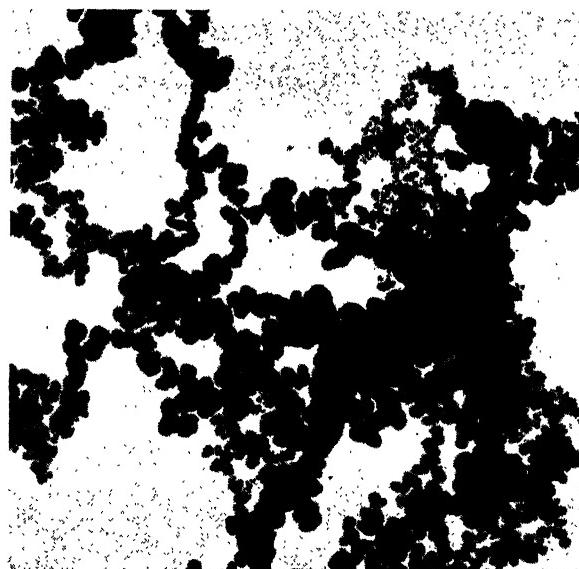
Figure 2. Semi-reinforcing furnace black after Cottrell agglomeration. (25,000x)

passing through the Cottrell precipitator. The average primary-particle size for this grade of black is about 0.07 micron. The tendency of the primary particles to form chain-like aggregates is very strikingly shown by these photographs, which were made by catching carbon-black smoke from a commercial unit directly on the resin-coated 2-mm diameter 200-mesh screen used in electron-microscope technique to hold the specimen. As far as is known, this has never been done before; all previously published photographs were made by redispersing collected carbon black by the dry-sparking or the lacquer-grinding technique. Interesting as these photomicrographs are, it is evident that the magnification is too great to give an overall measure of agglomerate size distribution.

Very soon after formation of carbon-black nuclei and growth of these nuclei to form the primary particles, or concurrently with the formation of primary particles, the particles form agglomerates,* ranging in size up to well above 10 microns. The left-hand panels in Figure 3, (a), (b), and (c) show photomicrographs of these agglomerates for several grades of furnace black. Agglomerates considerably larger than 5 to 10 microns may be formed when there is a sufficiently high initial concentration of primary carbon particles. Such large agglomerates are visible in the left-hand panel of Figure 3(a). Quantitative data on the size distribution of agglomerates in carbon-black smoke have not been published, as far as is known. No good method of obtaining such data has been developed, although there is considerable work under way now on this type of sampling problem.[5-7] The photomicrographs of agglomerates (shown in Figure 3) were made of samples caught on glass slides by quickly passing the slides through the smoke from commercial carbon-black units.†

In addition to a variable number of large (5 to 10 microns or more in diameter) agglomerates, there are many small agglomerates in a carbon-black smoke that has not been subjected to some external agglomerating force (Figure 3). These small agglomerates, which are a large number fraction of the total agglomerates, are less than about 5 to 10 microns in diameter, so that they settle very slowly and cannot be collected at high efficiency in cyclones. The weight fraction of agglomerates less than 5 microns in diameter seems to be a function of the total concentration of carbon black rather than a function of primary-particle size of the black. It is the agglomerates below 5 to 10 microns in diameter that present collection difficulties.

The range of effective agglomerate size is determined by inference from

---

* The problem of determining what is a *true agglomerate*, formed by chance collision of particles and agglomerates already formed, and what is a *structural* or *chain-like agglomerate*, formed integrally as a part of the carbon-black primary-particle growth process, is unsolved and is probably not pertinent to this discussion.

† Acknowledgments are due to Mary H. Martin and N. D. Steele, of Godfrey L. Cabot, Inc., for preparing these slides and photomicrographs.

(a)

(b)

(c)

(d)
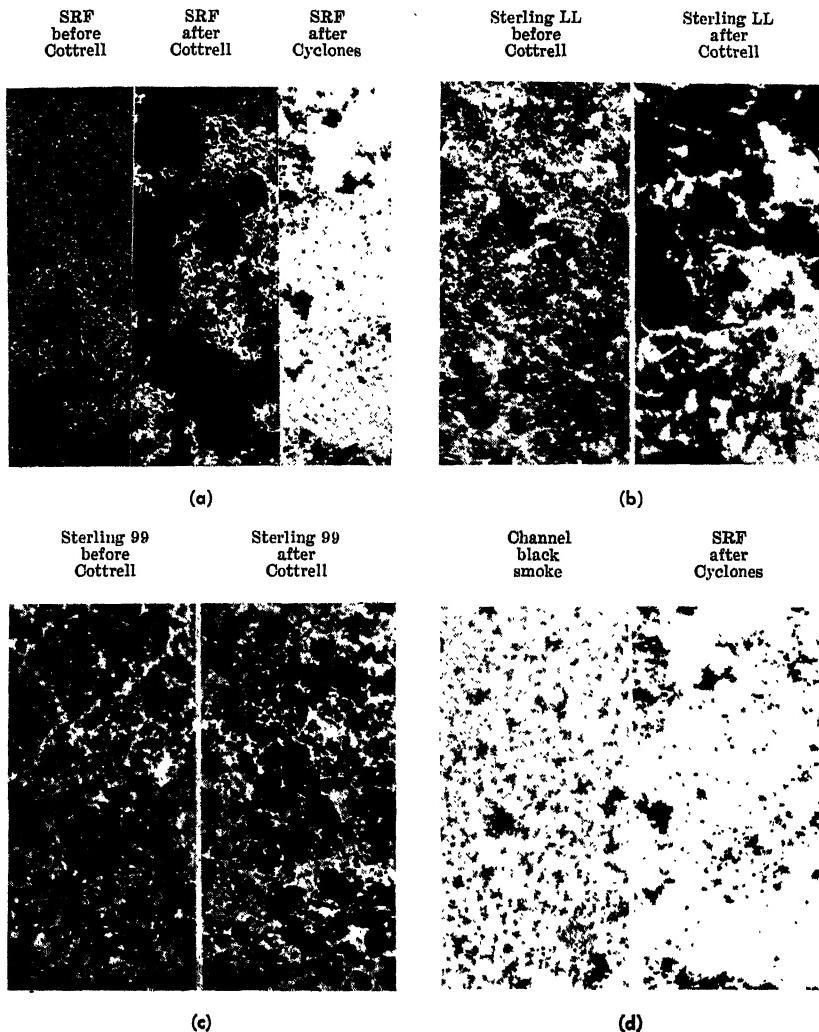
Figure 3. Agglomeration of several grades of furnace black. (100x)

the fact that, in most cases, carbon-black smokes settle only very slowly unless agglomerated by a superimposed force, such as prevails in electro-static precipitation, and from the fact that the agglomerates in carbon-black smokes which have not been subjected to superimposed agglomeration forces are visible in the ordinary light microscope. These two facts indicate that a large portion of the smoke consists of agglomerates in the range of 0.3 to about 5 to 10 microns. When a carbon black is formed under conditions that produce a high concentration of carbon particles in the smoke, particles may agglomerate rapidly enough to

allow practical collection in gravity settling chambers. This is the case
with lampblack made from creosote oil; the carbon black is formed just
above the surface of the vaporizing oil before the vapor is admixed to
any great degree with the combustion air. Some small agglomerates of
lampblack escape or are collected by secondary means, but they are a
minor weight fraction of the total black produced in this case.

Methods of Agglomerating and Collecting Aerosols. Aerosols formed
by condensation processes usually must be agglomerated before the dis-
persed material, whether liquid or solid, can be separated out in a prac-
tical way. On the other hand, the coarser particles of aerosols formed by
dispersion processes can usually be collected without a special agglomera-
tion step. Devices for recovering the dispersed particles of aerosols may
be divided into two groups according to their principal function:

    I. Agglomerating
       (a) Electrostatic precipitators
       (b) Sonic agglomerators
    II. Collecting and agglomerating
       (a) Gravity settling chambers
          (1) without contact surface
          (2) with contact surface
       (b) Inertial separators, dry or wet
          (1) centrifugal type
          (2) impingement type
       (c) Barrier devices (filters)
       (d) Thermal precipitators
       (e) Interphase transfer devices (scrubbers, using liquids
          as collection medium)

Devices having agglomeration as a principal function are usually fol-
lowed by those having collection as a principal function, although this
is often unnecessary in the case of electrostatic precipitators when operat-
ing on high specific-gravity solid aerosols or on liquid aerosols that col-
lect on the electrodes or agglomerate in the space between electrodes and
fall into the hopper without being re-entrained in the gas stream. Also
in the case of extremely dilute aerosols, an electrostatic precipitator may
function primarily as a collection device. In such a case, the particles
of the aerosol are moved (with little chance of agglomeration because of
the large spacing between particles) to the discharge electrode, where
they adhere to a dry or wet collector plate. An example of this is the
"Precipitron," which is used to remove dust from air in air-conditioning
installations. It is worth noting that in all the aerosol collection devices
agglomeration (cohesion) and adhesion play a very important part in the
proper functioning of the apparatus. For example, a bag filter will col-
lect fairly dilute carbon-black smoke even though the agglomerates are
much smaller than the pores in the filter cloth. The carbon agglomerates
adhere to the fibers and then other carbon agglomerates cohere to the

carbon adhering to the fiber. In this way a filter cake of carbon is built
up that is impermeable to even the smallest carbon agglomerate. If the
carbon-black smoke is too dilute, the carbon penetrates the fabric so
that it not only plugs it completely, but also destroys its flexibility to a
large degree.

There are a number of practical reasons why furnace-black aerosols
should be agglomerated before an attempt is made to collect them. It
would be beyond the scope of this paper to go into a discussion of all
the reasons. Suffice it to say that furnace-black collection is a two-step
process: agglomeration followed by collection.*

In view of this statement it may be somewhat puzzling that thermal
blacks are collected in bag filters. The answer is that thermal blacks are
produced in a cyclical process and are collected in an atmosphere con-
taining 80 per cent or higher concentration of hydrogen. Furthermore,
the pigment- and filler-grade thermal blacks (as contrasted to by-product
grades) are much more economical to handle when collected dry. A bag
filter is about the only practical device yet developed that will meet all
of the above conditions.

The electrical precipitator (Cottrell) functions very well for agglom-
erating furnace blacks and is used today in all furnace-black plants. How-
ever, it is a high first-cost apparatus and suffers somewhat from lack of
flexibility when switching from one grade of furnace black to another
and/or to different flow rates. It must be shut down for frequent clean-
ing of the electrical system and is sensitive to electrical surges due to
lightning storms or other causes. Power consumption for energizing the
electrodes and for pumping gas through the precipitator is very low.
The cost of the cyclones used in conjunction with electrostatic pre-
cipitators is relatively low compared to cost of the precipitator and
auxiliaries.

Because of the growing importance of furnace black, the increasing
difficulty of collecting finer particle-size blacks due to the low total black
concentration in these cases, and the great increase in construction and
equipment costs with a corresponding increase in natural gas value, it is
only logical that some effort would be made to develop agglomeration
apparatus of lower first cost and of equivalent or superior performance
compared with the electrostatic precipitator.

This paper presents the results of the first (as far as is known) fairly
large-scale experimentation, using actual commercial conditions, with
sonic energy as a means of agglomerating an aerosol, which in this case
was so-called *fine furnace black*. Because of the pioneering nature of the

* However, in at least one case a unit is being built to both agglomerate and collect
carbon black in an electrostatic precipitator. Collection efficiencies of 96 ± 2 per cent
at an inlet grain loading of 2.2 grains per cubic foot are confidently expected by the
builder.[s]
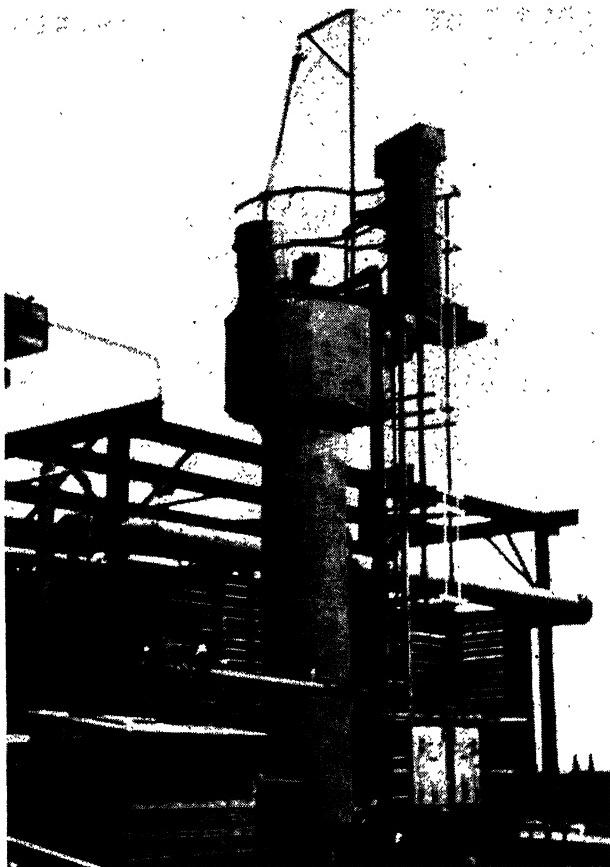
Figure 4. Sonic collection system for recovering sulfuric acid fog (center, foreground) installed at American Cyanamid Company, Warners Division, Linden, New Jersey.

work, a great many questions were left unanswered. Many of these questions have been partially answered by subsequent commercial-scale tests on other aerosols. It has been concluded, as a result of the experiments described here and subsequent commercial tests, that the next step in applying sonic agglomeration to furnace black might well be a full commercial-scale unit involving a gas flow of 30,000 cfm or higher. The data on sonic agglomeration presented later in this paper represent some of the first development work done by the Ultrasonic Corporation and should not be taken as indicative of the present state of the art, since considerable development has taken place since.*

* Commercial sonic agglomeration systems have been installed and are operating successfully to collect soda ash from paper mills and to recover sulfuric acid fog and fume. Figure 4 is a photograph of the installation at American Cyanamid Co. for the recovery of sulfuric acid fog.

Sonic Agglomeration Experiments

Sound Generators in Gaseous Media. Three types of apparatus are used to generate sound in a gas: (1) piston; (2) valve (siren); (3) gas current (whistle). The various types of sound generators have been reviewed by Sollner.[9]

The piston type includes crystals that vibrate by the piezo-electric effect, magnetostriction devices, and electromagnetic devices. Other methods of actuating a piston, e.g., a purely mechanical system for vibrating a piston or diaphragm, may be visualized. It is sufficient to say here that as yet no one has developed a piston device that is practical for genera-

$$f = N_R \cdot \left(\frac{REV}{SEC}\right)$$
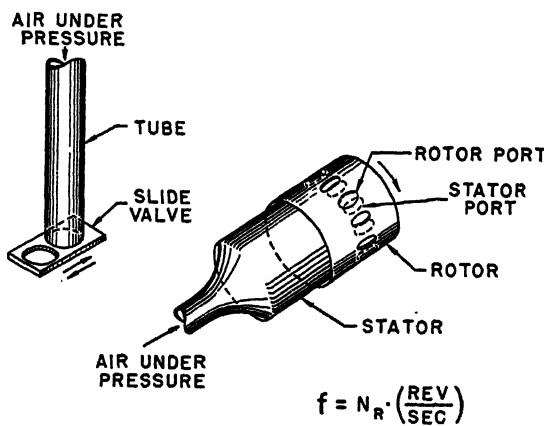
Figure 5. Schematic drawing of siren- or valve-type sound generator.

tion of high-intensity sound on a large scale in gases. St. Clair [10] used an electromagnetic device on a fairly large laboratory scale in studies on agglomeration of ammonium chloride and other smokes. A satisfactory piston-type sound generator for commercial use in gaseous systems would present an advantage over the valve type in that there would be no dilution of the gas being treated when using the piston-type generator.

In valve-type sound generators, a flowing gas stream is interrupted at a high frequency by a suitable valve arrangement. The simplest case would be that of a tube (left, Figure 5); the air discharge is interrupted at regular frequency by a slide valve. In order to get high frequencies a siren design (right, Figure 5) is employed commercially, in which there are N_S stator holes and N_R rotor holes. At a given speed of revolution (rps), the frequency of the sound is given by:

$$f = N_R \text{ (rps)}$$

The valve-type sound generator always involves dilution of the gas being treated by the gas passing through the valves of the generator. Use of

separating diaphragms between the sound source and acoustic chamber is not yet practical on a large scale.

In practice, a siren-type generator is driven by a variable-speed electric motor or a turbine. This type of generator can be driven by the rotor gas supply if the rotor and stator are designed properly. However, an external drive is needed for good control. The principal power input is the power required to compress the rotor gas. The rotor gas supply is at pressures up to about 5 or 6 lb/in ^{2} gauge. Higher pressures may be used, giving higher intensities of sound, but with existing siren-generator designs, above 5 or 6 lb/in ^{2} gauge, a very small gain in sound intensity is
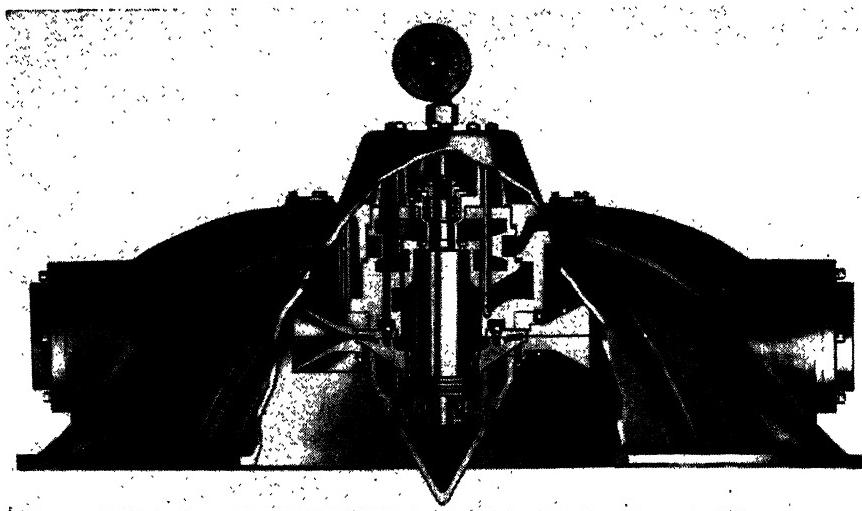
Figure 6. Cutaway view of the Ultrasonic Corporation U-2 siren-type sound generator.

at the expense of a large decrease in efficiency. At constant speed and rotor pressure, the total sound energy output is proportional to the total area of the rotor ports. Figure 6 shows a cutaway view of the Ultrasonic U-2 siren-type sound generator used for the work described in this paper. The problems of design and operation of siren-type generators are very similar to those encountered in high-speed turbine design and operation. Hence it may be concluded that such generators can easily be made practical commercial devices, a conclusion now borne out by commercial-scale operation in at least two cases.[11]

Gas current or whistle-type sound generators have been described in the literature.[12] These devices have not been developed to high enough efficiency to be of practical importance in commercial sonic agglomeration. An organ pipe is an example of a low-intensity whistle-type sound generator.

Effects of Sound Energy on Aerosols. Agglomeration of aerosols by sound energy has been studied more thoroughly with a view toward practical use than any other action of sound waves on colloidal systems. The state of knowledge up to 1944 has been well summed up by Sollner: [13] "Numerous attempts have been made to apply sonic coagulation methods to the precipitation of industrial fogs and smokes. Since coagulation occurs readily in streaming aerosols it is possible to handle large volumes; however, no definite statement can be made at present as to whether or not sonic coagulation may be developed into an industrial method; the chances that this may be achieved seem to be rather good."
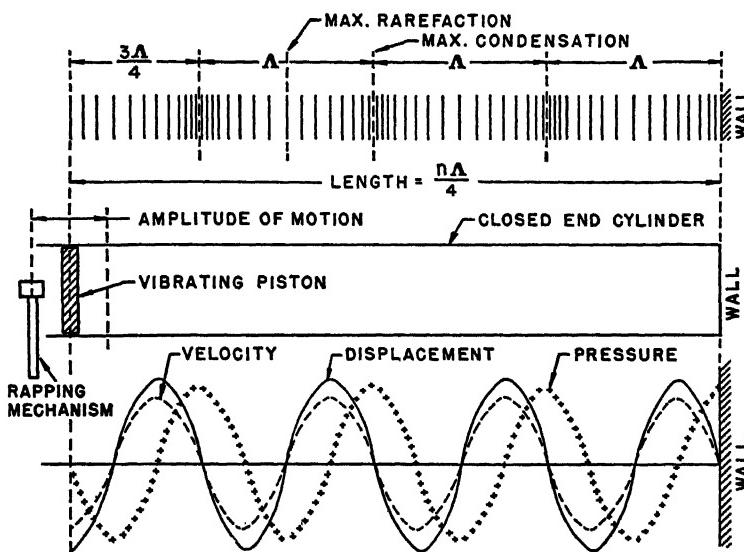
Figure 7. Sound effect in resonant system.

Up until 1946, when Horsley and associates of Ultrasonic Corporation developed their high-output siren, no practical device was available for generating high-intensity sound in gases on a large scale, so that prior to 1946 laboratory findings on sonic agglomeration could not be followed up by pilot-plant or commercial-scale tests. The theory of the agglomerating effect of sound is qualitatively well developed but quantitative calculations of agglomerating effect under commercial conditions cannot yet be made. It has not been definitely established which of the several mechanisms of agglomeration is most important under a given set of circumstances.

Figure 7 shows an idealized system in which the gas in a closed-end sound chamber is in resonant vibration with a source of sound waves,

in this case a vibrating piston-type sound generator. The length of the chamber is related, at resonance, to the wavelength of the sound by

$$L = \frac{n\lambda}{4}$$

where L is the chamber length, λ is the wavelength of the sound, and n is any odd integer. The top line of Figure 7 shows schematically the condensation and rarefaction of the gas at any instant, and the bottom line (the curve designated by crosses) shows the corresponding instantaneous pressure as a function of length. The points of maximum condensation and rarefaction will be reversed each cycle, but the points at which the pressure curve crosses the reference axis remain the same. The gas velocity and displacement curves are in phase with each other and are a quarter wavelength out of phase with the pressure curve. The points where the curves cross the neutral axis are called *nodes;* the maxima of the curves are called *antinodes* or *loops.*

At a given frequency of rapping, the energy input to the chamber remains constant as long as the amplitude of the piston is constant. At a given frequency, if the rapping is made more vigorous, the amplitude of the piston increases and the energy input to the chamber increases as the square of the amplitude. The energy input at constant amplitude increases as the square of the frequency.

It is clear that in the resonant system described there must be sound energy losses equal to the energy input at the piston or the amplitude of the piston would increase each time the sound wave is reflected from the end of the chamber back to the piston. In practice there are at least four ways in which the sound energy will be dissipated: (1) absorption at the wall and transmission through the wall; (2) damping due to viscous (frictional) effects in the gas; (3) losses through openings (in case of flow systems); and (4) absorption of energy in causing the aerosol particles to move (ultimately appears as a frictional effect in the gas).

Each time a sound wave strikes a rigid wall of smooth material (surface discontinuities small compared with the wavelength), the energy of the reflected wave is about 99 per cent of the energy of the incident wave, a loss on the order of 1 per cent [effect (1)]. About 10 per cent of this effect can be accounted for, as needed, to accelerate the molecules of the wall construction material, and about 90 per cent cannot be explained readily. Since this loss occurs each time the wave strikes the wall, the total loss by this mechanism can be quite appreciable. The second effect is small at low frequencies (on the order of 1000 to 10,000 cycles/sec) with low molecular-weight gases at ordinary pressures and temperatures. Effects (1) and (3) are the major effects in practical com-

mercial systems, i.e., flow systems. It may be possible to reduce effect (3) by installing acoustic filters of suitable design. Effect (4) is useful and is believed to be of appreciable magnitude, although no one has succeeded in measuring it with high precision.

The intensity of sound energy is measured by

$$\frac{energy}{(unit\ time)\ (unit\ area)}$$

in the direction perpendicular to the travel of the wave. A common unit of intensity is watts/cm² or

$$\frac{ergs}{(sec)\,(cm)^2} \times 10^7$$

The decibel notation is also used to designate intensity. The number of decibels between two intensity levels is 10 times the log of the intensity ratio:

$$N = 10 \log \frac{I_i}{I_o}$$

where N is the decibel rating corresponding to the intensity I_i. On this scale, the faintest sound of frequency 1000 cycles/sec discernible by the normal human ear corresponds to an absolute energy intensity of 10^{-16} watts/cm² and to a pressure effect of 0.0002 dyne/cm². Thus, for an intensity of 160 decibels, the absolute intensity is:

$$I_i = I_o \left(antilog\ \frac{N}{10} \right) = 10^{-16} \cdot 10^{16} = 1\ watt/cm^2$$

Another term of significance in acoustic systems is the energy density, usually expressed in ergs per cm³.

Brandt, Freund, and Hiedemann [14-16] have studied, by means of high-speed photomicrography, the motion of aerosol particles in a sound field and have published photomicrographs of the aerosol before and after agglomeration. These same authors [17-22] and others [23-27] have discussed the movement of suspended particles in a sound field and have developed mathematical relationships to describe particle motion. Recently St. Clair [28-30] has published the results of a series of experiments on sonic agglomeration of ammonium chloride smokes. He shows by photographic evidence (Figure 8) that the agglomerated aerosol collects in regions spaced ½ wavelength apart, corresponding to nodes or antinodes, and concludes that the radiation-pressure effect is of controlling importance in causing agglomeration.

It is recognized that at least three mechanisms are operative in causing agglomeration of an aerosol in a sound field: (1) Covibration of particles in the vibrating gas as a result of the drag between particles and dispersion medium; this causes an increased rate of particle collision. (2) The Bernoulli effect due to the constriction or expansion of the dis-

persion medium as it moves rapidly back and forth between or parallel
to closely-spaced particles. (3) The radiation-pressure effect, arising be-
cause of difference in momentum of the dispersion medium on opposite
sides of the particle.
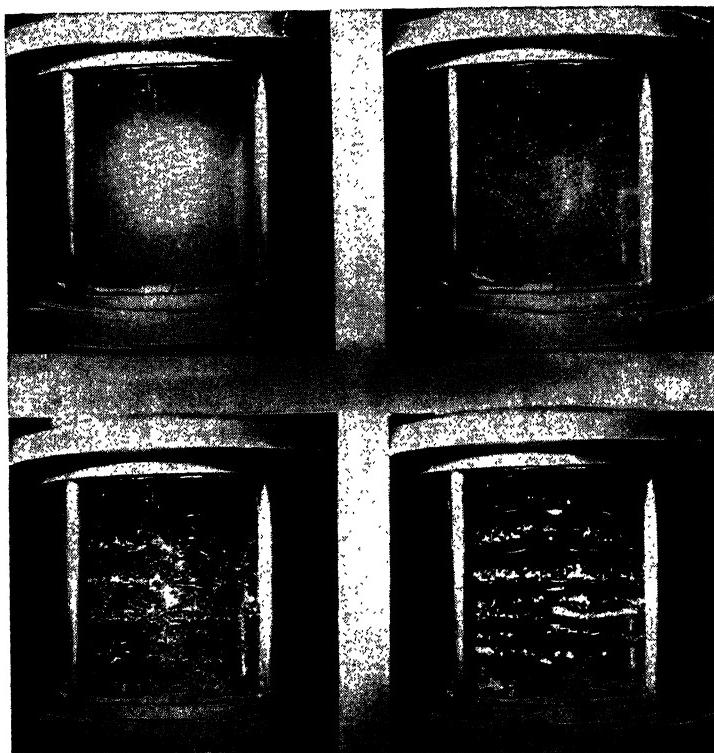
Figure 8. Concentration of agglomerated aerosol (ammonium chloride) at
nodes and antinodes. (From St. Clair [28-30].)

If a stable aerosol * is passed through a sound field of high intensity,
the aerosol particles are moved back and forth at the frequency of the

* In a carbon-black aerosol, during and immediately after formation, the particles
are subject to Brownian motion (which is greatly increased by the high temperature
of the furnace process), which brings about a rapid rate of collision between particles
until the agglomerates so formed either become too large to be affected by the impacts
of the molecules of dispersion medium or become too far apart for frequent collision.
When this stage is reached, the aerosol is relatively stable and will settle only very
slowly under the influence of gravity. Channel-black smoke, for example, will remain
suspended on the horizon for days at a time when there is little turbulence in the
air below a few thousand feet above the ground. The particles of a stable aerosol
may be "protected" from agglomeration by a repulsion effect if all particles bear
electrical charges of the same sign. The particles or the agglomerates are surrounded
by a stagnant layer of gas that may tend to reduce the number or the effectiveness
of the collisions between particles and agglomerates. The adsorption of gas in a thin
layer on the surface of the particles is thought to make agglomeration more difficult.

sound and at a velocity varying with the size of the particle, but always less than the velocity of motion of the dispersion medium. [This is mechanism (1) described above.] The largest particles move very little because of their high ratio of mass to drag; the smallest particles move almost at the velocity of the dispersion medium. The higher the intensity of the sound, the greater the rate of acceleration of the particles and hence the greater the energy of the impact between particles. For each size of particle there is an optimum frequency of sound, at a given sound intensity, that will produce the greatest degree of motion of the particle. However, because of the broad size distribution of particles and agglomerates, the range of optimum frequency may be rather broad. It is quite probable that this motion effect is of controlling importance in the agglomeration of carbon-black aerosols when subjected to a sound field for only a few seconds, as would be the case in any commercially practical apparatus. It does not seem likely that the radiation-pressure effect described below would be the major cause of agglomeration when the contact time is low (less than 2 or 3 sec).

The Bernoulli effect is of importance only when particles are quite close together. When the line of centers of the particles is at right angles to the sound field, the particles are attracted. When the particles are arranged with their line of centers parallel to the sound field, the particles are repelled.

The radiation pressure in a sound field may be measured by suitable delicate instruments. Its magnitude may also be calculated.[31] Radiation pressure is demonstrated by the fact that a standing wave sound field in a gaseous medium will support objects of considerable size, e.g., a ping-pong ball or a fifty-cent piece.

Laboratory Apparatus and Procedure. The first experiments were made in a "Pyrex" glass jar into which sound was transmitted through a stretched rubber diaphragm about 1 mil in thickness. This jar, which held about one-half cubic foot of smoke, was flushed with carbon-black smoke until all the original air was displaced. The inlet and outlet were then closed and sound directed against the diaphragm. Agglomeration was followed qualitatively by changes in intensity of transmitted light measured with a photocell. All measurements were relative as no calibration of light transmission versus agglomerate size could be made. Studies of light transmission versus time were made for carbon-black smoke, without sound and with sound of varying frequencies and intensities. No continuous-flow experiments were made.

From the start, the laboratory experimenters were beset by many problems. The principal difficulty was the precipitation of carbon black on the walls of the "Pyrex" glass jar, which interfered with the light-transmission readings. This difficulty can be overcome in a flow system

by suitable means, but in a batch system it can only be minimized, not eliminated, by keeping the smoke temperature the same as the temperature of the glass, thus avoiding the thermal-precipitation effect.

The carbon-black smoke was made by burning natural gas; to avoid condensation of water vapor on the chamber walls, the smoke was passed through a "Dry-Ice" trap to remove water vapor. This meant that there probably was a fog of water droplets mixed with the carbon-black particles, making agglomeration of the carbon black easier.* Hence the laboratory experiments did not closely duplicate expected commercial conditions. The rubber diaphragm made experiments at temperatures above the smoke dew-point impossible even if wall condensation could have been avoided.
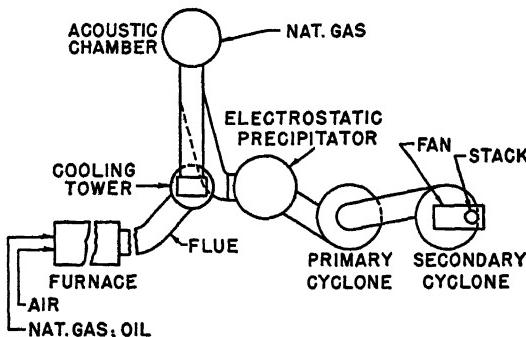
Figure 9. Flow diagram for sonic agglomeration experiments.

Sound intensity was measured in the empty chamber using a microphone device of conventional construction. The precision of this measurement is low. It is interesting to note that a microphone makes use of the piezo-electric effect to measure sound intensity, and it will be recalled that the piezo-electric effect can also serve as a source of sound, e.g., quartz crystal sonic generators. Sound was generated by an Ultrasonic Corporation U-1 siren-type generator driven by compressed air. This model of generator is now obsolete.

It was concluded that the laboratory experiments were of little value in judging the practicability of commercial sonic agglomeration of carbon black, although these experiments did demonstrate the pronounced agglomerating effect of sound at the 150-decibel level. Hence it was decided to go to pilot-plant scale in further experimentation.

Pilot-Plant Setup.† For pilot-plant work, a sonic agglomerating system

* Experiments on agglomeration of artificial water fog showed that fog agglomerated more readily than carbon-black smoke under comparable conditions. Theory indicates that water fog mixed with the smoke would greatly assist the agglomeration of the latter.

† W. L. Loving and N. D. Steele, of Cabot Carbon Company, Pampa, Texas, were in charge of the pilot-plant work. Ultrasonic Corporation, Cambridge, Massachusetts,

was installed in parallel with a Research Corporation electrostatic precipitator (Cottrell) in an existing furnace-black pilot plant. The layout of the apparatus is shown in Figure 9, and Figure 10 is a photograph of the apparatus. Figure 11 shows a commercial Cottrell cyclone system of 50,000-cfm rated capacity. The similarity of pilot-plant and
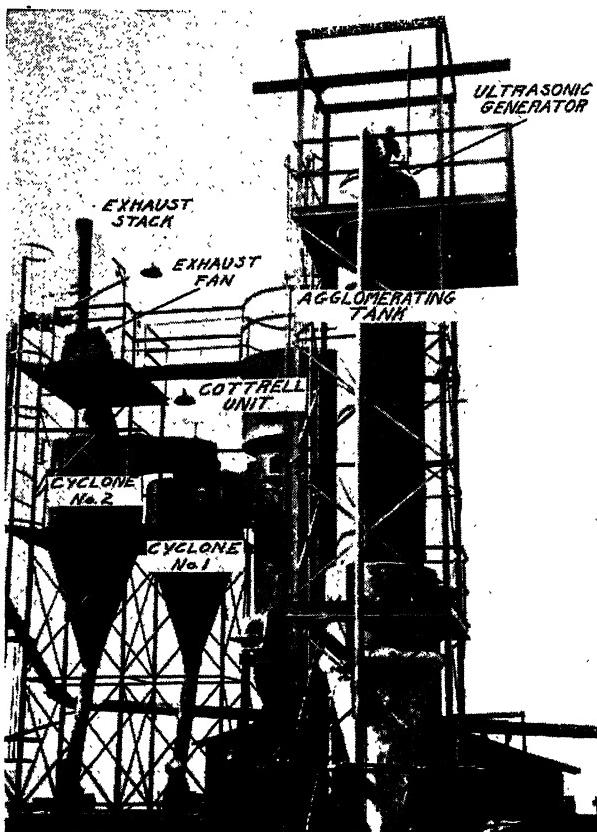
Figure 10. Pilot-plant installation (general equipment) for sonic agglomeration experiments.

commercial apparatus is evident. The rated capacity of the pilot-plant Cottrell-cyclone system is 3500 cfm at flow conditions, although this rated capacity has very little significance since it must necessarily be based on some arbitrary degree of collection of a certain type of black

at a specified range of inlet grain loadings.* Suffice it to say that the rated-capacity figure reflects the flow rates (contact times) and grain loadings encountered in commercial precipitators on furnace black. Since the smoke passed in series through the sound chamber and the Cottrell, it was possible to test one or the other or the two in series for agglomerating the smoke produced under any given set of furnace conditions.

Fine furnace-blacks were made in the furnace, using commercial burners and commercial fuel-air ratios, so that the smoke formed was accurately representative of commercial conditions. The primary-particle
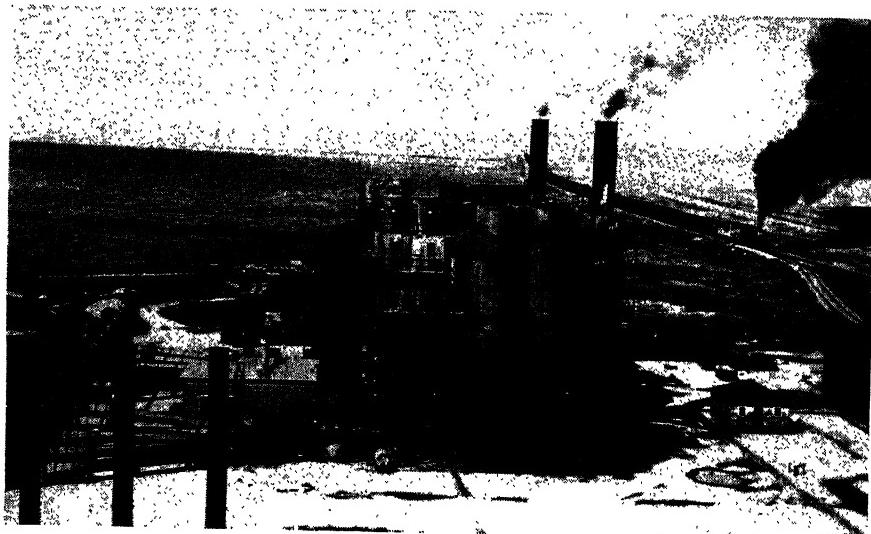
Figure 11. A commercial Cottrell-cyclone collection system of 50,000-cfm rated capacity.

size and grain loading of the carbon-black smoke could be varied more or less independently over a rather limited range by varying the degree of oil-enrichment of the natural gas serving as the primary fuel. The smoke temperature could be adjusted by water sprays in the cooling tower down to about 100°F above the dew point. Lower temperatures caused condensation difficulties in the agglomeration and collection system. The rate of smoke delivery to the sound chamber was about 1000 to 1200 cfm at chamber conditions.

The sound generator was the Ultrasonic Corporation U-2, driven by a turbine powered with high-pressure natural gas. Natural gas was used

* This particular Cottrell-cyclone system is rated to collect 90 per cent or more of semi-reinforcing furnace black at a flow of 3500 cfm (450°F, 1 atm pressure). The inlet grain loading for this type of black is approximately 5 grains per cu ft at flow conditions.

as the rotor fluid in amounts averaging about 1000 cfm measured at 14.7 psia and 60°F. Thus the dilution of the smoke by the rotor gas was about 1 to 1, which presented a serious disadvantage to obtaining data truly representative of expected large-scale commercial conditions (volume-flow rates of 20,000 to 65,000 cfm at 450°F and atmospheric pressure are common on a commercial scale). The smoke entered the sound chamber near the top and left near the bottom. The sound generator was mounted at the top of the chamber (Figure 12). Very little black collected in the sound chamber. Most of the agglomerated black was swept along with the gas stream and collected in the two cyclones. It was interesting to observe that no carbon collected on the walls of

Figure 12. Sonic agglomeration chamber.
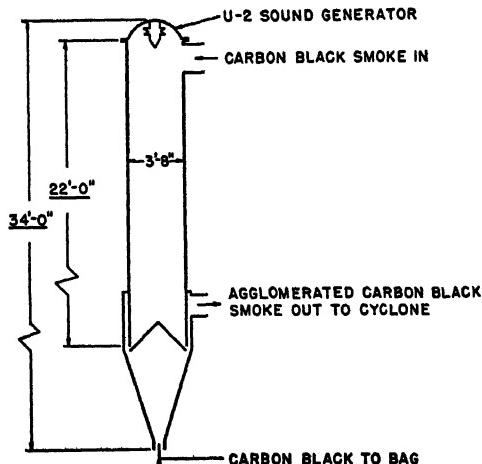
the chamber, as was expected. Apparently the vibration kept the surface clean. Grain-loading tests by Research Corporation method were made at agglomerator inlet and stack outlet on most runs. The total black collected in sound chamber, precipitator, and cyclones was accurately weighed on each run. Collection efficiency could be calculated either from the stack loss or from the difference between the known carbon input to the agglomerator and the black collected. For a given set of furnace conditions the carbon yield was accurately known from previous experiments. The flows of natural gas, air, and oil to the furnace were measured, the gas and air on recording meters and the oil by gauging tanks. The water-vapor content of the smoke stream entering the agglomerator was measured by condensing the water vapor down to a given dew point in a metered volume of gas. The water collected plus that calculated to be in the saturated gas gave the total water content. The rotor gas supply and generator *rpm* were measured. The turbine gas supply was not metered, but the inlet pressure was

recorded. A previous calibration gave the relation of gas flow through the turbine and *rpm* as a function of inlet pressure.

Sound-intensity measurements were made with a microphone when no smoke was passing through the chambers; the microphone could not be used in the smoke because of the high temperature. The sound-intensity measurements were not traverses but measurements of average maximum intensity. Later experience in sound measurement has indicated that the overall average intensity corresponding to the average maximum readings obtained would have been about 155 decibels, and this figure is used with the data in this paper since it is the best estimate available. If the intensity could have been measured with smoke in the chamber, it would have been found lower than in the empty chamber because of the absorption of sound energy in moving the aerosol particles.

Theoretical power input could be calculated as that available from expanding turbine and rotor gas reversibly down to the sound-chamber pressure (13.2 psi atmospheric pressure). No method has been devised as yet to measure the acoustic power output of a siren-type sound generator in an enclosed field. Hence there is no way of computing the efficiency of conversion to acoustic energy of the total energy input in the system. Measurements made in a free field indicated efficiencies for this type generator in the range of 10 to 50 per cent, depending upon operating conditions and design. The agglomerated and unagglomerated smoke were sampled during one run so that photomicrographs could be made.

Summing up the significance of the measurements that could be made with the pilot-plant apparatus: Carbon-black smokes of a limited range of primary-particle size and grain loading could be subjected to sound intensities of not greater than about 155 decibels for a limited range of contact times. Dilution of the smoke by rotor gas was about 1 to 1. The degree of collection of the sonically agglomerated smoke in a cyclone system of arbitrary design (not the design for optimum collection) could be measured, and observation of the smoke itself could be made to discover qualitatively the degree of agglomeration.

Laboratory Data. The most significant laboratory data are summarized in Figure 13. These data show the effects of sound of two different intensities on clearing carbon-black smoke. For comparison, a blank run was made in which carbon-black smoke was allowed to stand without sound treatment. The curve for agglomeration of water fog is also shown in Figure 13.

An attempt was made in the laboratory experiments to find the optimum sound frequency for carbon-black agglomeration, but the results were inconclusive and the data are not presented here. It was estimated from the laboratory data on carbon-black smoke and artificial

water fog that the frequency range of 3000 to 4000 cycles/sec would be optimum for carbon black.

The laboratory data served merely to show that sound energy at intensities in the range of 145 to 150 decibels and at frequencies of the order of 3000 cycles had a pronounced agglomerating effect on carbon-black smoke that had been dehumidified and cooled to near room temperature. Very rough tests gave the disturbing indication that smoke which had not been passed through a "Dry Ice" trap for cooling and dehumidification would not agglomerate as well as smoke that had been so treated, but the practical significance of this difference could not be judged.
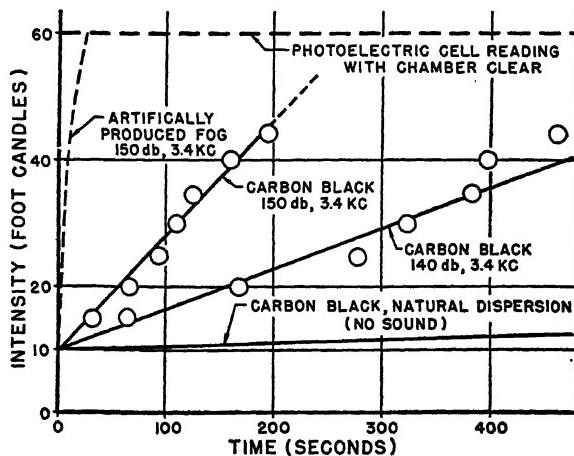
Figure 13. Transmitted light intensity versus time (measurement of dispersion in test chamber), showing the effects of sound of two different intensities on clearing carbon-black smoke. (The curve for agglomeration of water fog is also shown.)

Pilot Plant Results. Because of the pioneering nature of the work, the experimental setup as well as the data obtained were both inadequate for application to commercial design. Many of the variables could be measured only with limited precision. Grain-loading tests were particularly tedious and limited in precision. Some of the variables could not be measured at all; for example, the efficiency of conversion of the energy in compressed natural gas to sound energy. On the other hand, such trends as could be observed can be relied upon beyond all reasonable doubt, since enough data were obtained to give consistent and reasonable curves. Because the quantitative data are of limited precision and extent, it is felt that graphical presentation of the data is adequate. The detailed data would require space that can be used to better advantage for discussion.

Figure 14 shows a series of runs made to determine the optimum sound frequency for carbon-black agglomeration. While these data substantiate the laboratory data in showing an optimum frequency in the range of 3000 to 5000 cycles/sec, it is believed that the frequency effect

FINE FURNACE BLACK, 92 SCALE, 2 GR / FT³

CARBON BLACK RECOVERY VS. FREQUENCY
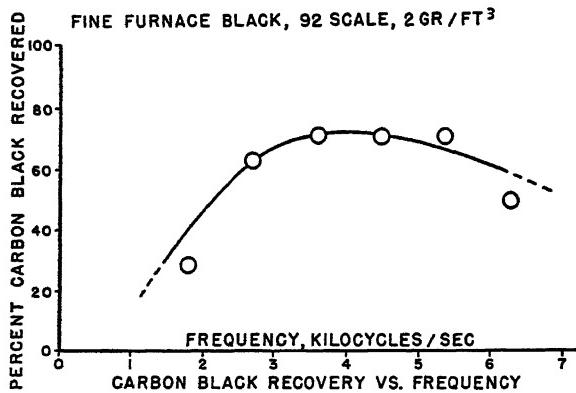
Figure 14. Carbon-black recovery versus frequency, for fine furnace-black at a sound intensity of approximately 155 decibels.
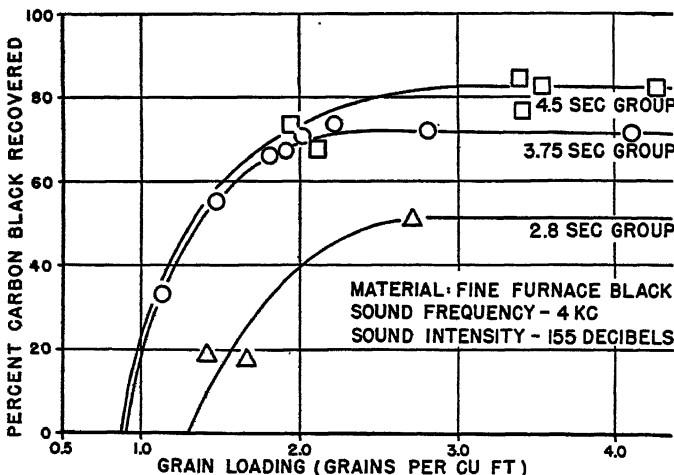
Figure 15. Carbon-black recovery versus carbon-black concentration.

is complicated by the characteristics of the sound generator. The optimum frequency may be as much due to the higher efficiency and higher sound intensity of the generator at certain frequencies as to the existence of a true optimum frequency range for agglomeration in the ranges studied.

Figure 15 shows the best of the basic data on collection of a given type of carbon black versus grain-loading and contact time, frequency and

intensity remaining constant. These data could be roughly verified by an experienced observer looking at the stack, and this verification served as a check against gross errors in collection-efficiency measurements.

An interesting observation was made by quickly dipping a glass slide into the smoke entering the sound chamber and comparing this slide at a magnification of 100 with one obtained in a similar manner on the agglomerated smoke at the sound-chamber outlet. The change in degree of agglomeration is very striking. In the agglomerated smoke the small agglomerates have been almost completely eliminated. The method of
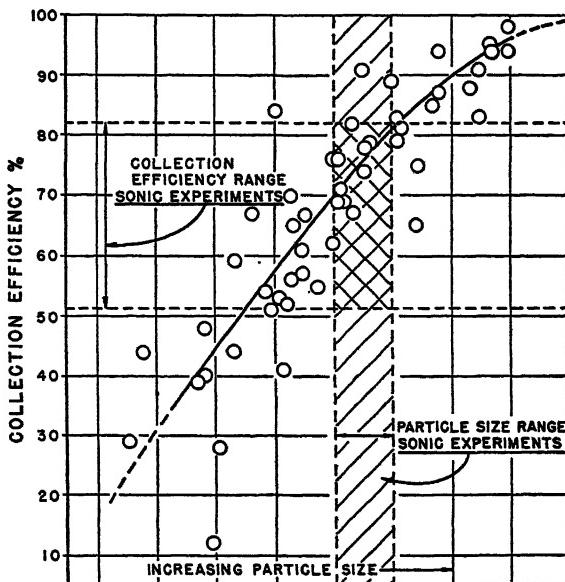
Figure 16. Collection characteristics for 3500-cfm pilot-plant Cottrell-cyclone system.

sampling is open to some question in that the larger agglomerates would be caught more easily than the smaller ones. However, in view of the relatively low velocity of the smoke and the rapid insertion and removal (in 2 or 3 sec) of the slide normal to the smoke flow, it would seem that a representative sample would be obtained. The slides were examined carefully over the entire area of deposited smoke to get a representative field for photographing.

No mechanical difficulties of any consequence were experienced with the sound generator, but no extended runs (over about eight hours) were made. The noise level was not great and no strong complaints were voiced by the operators. The high-pitched whine seemed to have a slightly annoying psychological effect. The chamber was not insulated

or otherwise specially equipped to cut down the intensity of the escaping sound.

Comparison Between Sonic and Cottrell Collection Systems. After the sonic-collection work was completed, a Pangborn leaf-type bag filter with asbestos bags was installed in the furnace-black pilot plant used for the sonic tests. This bag filter was used in a large number of subsequent runs to measure the efficiency of the Cottrell-cyclone system when making blacks of the same type as used for the sonic-agglomeration study.
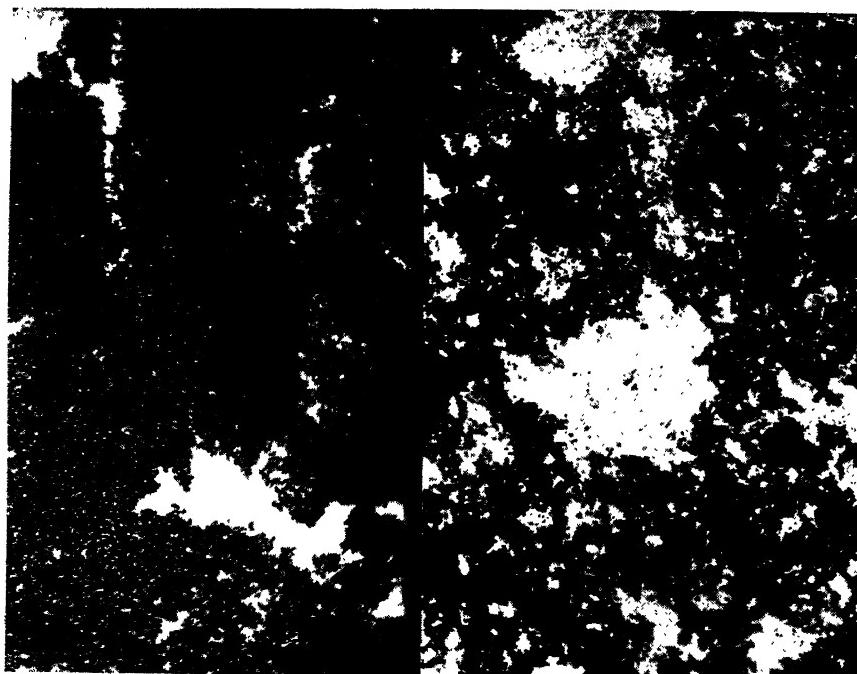
Figure 17. Photomicrograph of agglomerates of fine furnace black of primary particle size 0.04 to 0.05 micron. *Left:* Before sonic agglomeration; *Right:* After sonic agglomeration. (100x)

The bag filter gave substantially complete collection of the carbon in the exhaust from the secondary cyclone. Figure 16 shows the collection-efficiency data so obtained. The flow rates used in the subsequent runs with the bag filter were in the range of 1000 to 2000 cfm at flow conditions, which overlaps the range of the sonic experiments. It may be explained at this point that in practically all runs represented in Figure 16, the inlet grain-loading is lower, the lower the particle size of the carbon black. Those points that lie more than about 10 per cent below or above the curve, for the most part represent cases in which (1) the grain loading is abnormally low or high because of freak conditions, or

(2) the flow rate is abnormally high (above 2000 cfm) or low (below 1000 cfm).

At the lower and upper limits of the particle-size range covered in the sonic experiments, the Cottrell-cyclone system showed collection efficiencies of 60 to 80 per cent and 70 to 90 per cent respectively, while the sonic experiments covered an efficiency range of 52 to 83 per cent. No direct comparisons could be made between the sonic system and Cottrell system, as the only factor that could be held constant in operating the two systems was the use of the same cyclones for collecting agglomerates. However, it is fair to say that the sonic system showed efficiencies in the same range as the Cottrell system when both were operated over the same ranges of flow, particle size, and grain loading. The range of collection efficiencies for the sonic system lies somewhat lower because of the excessive dilution from rotor gas. Most of the sonic runs were in the lower grain-loading range (1½ to 2½ grains/cu ft) and higher flow ranges (2000 cfm), while most of the Cottrell runs were in the higher (2½ or greater) grain-loading, and lower (1000 cfm) flow range. It is concluded that, within the limits of the data, the sonic collection system gave performance equivalent to that of the Cottrell system.

Figure 17 shows a fine particle-size furnace black before (right) and after (left) passing through the sound field. The change in state of agglomeration is striking. Figure 3(c) shows the same comparison for a slightly larger particle-size furnace black before and after passing through a commercial (50,000-cfm) Cottrell agglomerator. It happens that this commercial grade of black (fine furnace-black, Sterling 99, mean particle size about 45 millimicrons) is particularly hard to collect in the conventional Cottrell-cyclone system. The photomicrograph suggests that this is partly because of poor agglomeration. Figure 3, (a) and (b), shows the greater degree of agglomeration that is obtained with the Cottrell agglomerator on still coarser grades of furnace black (about 70 millimicrons and 55 millimicrons mean diameter respectively). The photographic evidence suggests that a suitable photographic technique might be developed for study of the effectiveness of sonic versus Cottrell agglomeration. This kind of study is now being made under carefully controlled laboratory and pilot-plant conditions.

Commercial Implications. Sonic collection of furnace carbon-black could be employed at any time with full assurance of collection efficiencies as good as with electrostatic agglomeration systems. The sonic apparatus is simple and dependable. In this respect it is similar to a high-speed pump or turbine. The only question yet to be answered is the total operating cost and first cost for a given case. Further experimental data and correlations are required to put sonic collection on a firm design basis.

Conclusions

(1) Sonic energy at the intensity level of about 155 decibels for 2½ to 5 seconds contact time appears to be at least equivalent, in agglomerating effect on fine furnace-black, to a pilot-plant electrostatic precipitator which duplicates commercial precipitator performance and operates at commercial voltages.

(2) The effectiveness of sonic agglomeration is a function of the product of contact time and sound intensity up to a limiting contact time, after which little further agglomeration takes place. Improvement in sound generator and chamber design to give higher intensities at no sacrifice of energy efficiency may be expected to decrease significantly the contact time necessary for a given degree of agglomeration.

(3) There is no indication that sonic agglomeration, followed by dry cyclone collection, will give substantially complete collection of furnace carbon black (residual concentration at flow conditions of about 0.05 grain per cu ft or less). This system accomplishes a degree of collection of furnace carbon back comparable to that in a Cottrell dry cyclone system and not comparable to that obtained with a bag filter or wet scrubber.

(4) The investment and operating costs of sonic agglomeration of furnace blacks have not yet been determined. The final answer will have to be obtained on commercial-scale apparatus or on apparatus which consists of a unit identical with the several units of a multi-unit commercial apparatus. Two important efficiencies are involved: the efficiency at which shaft energy can be converted into sonic energy at the required intensity level; and the efficiency of utilization of sound energy for particle agglomeration. These two efficiencies cannot be studied separately in a practical commercial system, at least not at the present state of our knowledge, because both are closely related to chamber design.

REFERENCES

1. Perry, J. H., "Chemical Engineers' Handbook," 2nd ed., pp. 1850-84, New York, McGraw-Hill Book Co., Inc., 1941.
2. Gibbs, W. E., "Clouds and Smokes," pp. 57-86, London, J. and A. Churchill, Ltd., 1924; also his paper "Aerosols" in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. I, pp. 393-419, New York, Chemical Catalog Co. (Reinhold Publishing Corp.), 1926.
3. "Disperse Systems in Gases; Dust, Smoke and Fog," *Trans. Faraday Soc.*, **33**, 1042 (1936).
4. Newbury, W. C. (Samuel Cabot, Inc., Boston, Mass.), Personal communication.
5. Sonkin, L. S., *J. Ind. Hyg. Toxicol.*, **28**, 269 (1946).
6. May, K. R., *J. Sci. Instruments*, **22**, 187 (1945).
7. Johnstone, H. F., and Roberts, M. H., *Ind. Eng. Chem.*, **41**, 2417 (1949).
8. Linthicum, F. H. (Koppers Co., Inc.), Personal communication.

9. Sollner, K., "Sonic and Ultrasonic Waves in Colloid Chemistry," in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. V, pp. 338-46, New York, Reinhold Publishing Corp., 1944.

10. St. Clair, H. W., *Rev. Sci. Instruments,* 12, 250 (1940).

11. Neumann, E. P. (Ultrasonic Corp., Cambridge, Mass.), Personal communication.

12. Hartmann, J., "Modus Operandi of the Air-jet Pulsator" (Vorträge auf d. Geb. d. Aerodynamik u. verw. Geb. Aachen), Berlin, J. Springer, 1929.

13. Sollner, K., "Sonic and Ultrasonic Waves in Colloid Chemistry," in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. V, p. 363, New York. Reinhold Publishing Corp., 1944.

14. Brandt, O., Freund, H., and Hiedemann, E. *Z. Physik,* 104, 511 (1937).

15. Hiedemann, E., "Grundlagen und Ergebnisse der Ultraschallforschung," Berlin, W. de Gruyter & Co., 1939.

16. Brandt, O., and Hiedemann, E., *Trans. Faraday Soc.,* 32, 1101 (1936).

17. Brandt, O., *Kolloid-Z.,* 81, 2 (1937).

18. Brandt, O., and Freund, H., *Blätter Untersuch.- u. Forsch.- Instrumente,* 9, 57 (1935).

19. ———, and ———, *Z. physik, chem. Unterricht,* 48, 259 (1935).

20. ———, and ———, *Z Physik,* 95, 415 (1935).

21. Hiedemann, E., and Brandt, O., *Kolloid-Z.,* 75, 129 (1936).

22. Brandt, O., Hiedemann, E., and Freund, H., *Kolloid-Z.,* 77, 103 (1936).

23. König, W., *Ann. Physik.,* 42, 353, 549 (1891).

24. Lewis, E. P., and Farris, L. P., *Phys. Rev.,* 6, 491 (1915).

25. Wagenschein, M., *Ann. Physik.,* 65, (4), 461 (1921).

26. Laidler, T. J., and Richardson, E. G., *J. Acoust. Soc. Am.,* 9, 217 (1938).

27. Mache, H., *Meteorol. Z.,* 50, 393 (1933).

28. St. Clair, H. W., Spendlove, M. J., and Potter, E. V., "Flocculation of Aerosols by Intense High-Frequency Sound," *U. S. Bur. Mines, Rept. Invest.,* No. 4218 (March, 1948).

29. St. Clair, H. W., *U. S. Bur. Mines, Rept. Invest.,* No. 3400, 51 (1938).

30. ———, *Conf. Met. Research, Met. Div., Bur. Mines* (Salt Lake City), 99 (May, 1940).

31. King, L. V., *Proc. Roy. Soc. (London),* 147A, 233-6 (1934).

ORGANOPHILIC CLAYS

Ernst A. Hauser

*Professor of Colloid Chemistry, Mass. Institute of Technology, Cambridge, Mass.
and Worcester Polytechnic Institute, Worcester, Mass.*

Colloidal Clays

During the last two decades, considerable progress has been made in research pertaining to the composition, structure, and colloidal properties of clay minerals.[6, 7] Only recently, however, has it been shown that the combination of clay minerals with certain organic compounds forms an interesting bridge between the inorganic and organic kingdoms of matter.[7] This being the case, it might aid future research to apply some of the terminology used in high-molecular organic colloid chemistry to investigations pertaining to colloidal clays.

The basic constituents of all clays are the silicon tetrahedron, the magnesium octahedron, and the aluminum octahedron; they are the monomers. By condensation polymerization, hydrated silica and the minerals brucite and gibbsite, respectively, are formed. Occasionally a copolymer of magnesium- and aluminum octadedra is formed, which carries in its lattice an unbalanced structure of pronounced net negative charge. On copolymerizing by condensation one hydrated silica layer and a gibbsite layer, the clay mineral kaolinite is formed. If a brucite layer is copolymerized with two hydrated silica polymers, the clay mineral talc is formed. Copolymerization of two hydrated silica polymers with a hydrated magnesium- and aluminum octahedron copolymer results in the clay mineral known as substituted montmorillonite or bentonite.[8] This clay, having a very unbalanced lattice structure, can also be classified as a colloidal electrolyte.[17] It is composed of an anion of colloidal dimensions with cations adsorbed to it from the dispersion medium or by dissociation into the dispersion medium of ions loosely adsorbed in its structure.

The most important and interesting phenomenon exhibited by those clay minerals which will, in contact with water, form a colloidal micelle is the ion-exchange reaction. This reaction was originally recognized by

431

Thompson in his investigation on soil properties, but the first systematic work on the subject was done by Way.[24]

The ultimate clay crystal carries a net negative charge, the result of either anion adsorption onto its surface, or an unbalanced crystal lattice. Whatever the basic cause may be, one can picture the individual ultimate clay particle as a very complex anion. To balance its charge, the particle tends to adsorb the necessary number of cations available in its environment. When such a clay particle is dispersed in water, these cations will hydrate and, depending on their valency and degree of hydratability, dissociate to a certain distance from the surface of the particle, thereby building up a diffuse electric double layer and forming a colloidal micelle. One may compare the suspended clay particle with a dissociated electrolyte, since the size of one of its ions falls within the colloidal range of dimensions. A similar condition exists in the case of soap, where the sodium ion, owing to its hydration, will dissociate from the fatty-acid ion when the soap is brought into contact with soft water. If hard water is used, however, the divalent and less hydrated calcium ions contained therein will exchange for the sodium ions and form the far less soluble calcium soap.

By exactly the same mechanism, the counter ions of the clay particle can be exchanged with ions from the dispersion medium, if the resulting micelle will have less tendency to hydrate and/or carry a lower charge.[6] This reaction is, of course, more pronounced the more ions of a high degree of hydration are present in the clay under investigation. The ability to adsorb cations, therefore, depends on the structural configuration of the nucleus of the colloidal-clay micelle. Many of the properties of colloidal clays such as plasticity,[9] dry strength, thixotropy,[10] dilatancy,[11] and rheopexy,[12] could be satisfactorily explained [6] on the basis of this behavior. All these theories, however, are founded on the hydrophilic properties of the surface of the clay micelle and the inorganic ions involved in the exchange reactions.

Organophilic Clays

The first reactions between bentonites and organic bases were found to be principally of the base-exchange type. If bentonite is treated with an organic base like amylamine, or diamylamine, the reaction leads to partial or complete saturation of the silicate. The less soluble the formed silicate complex is, the more possible it is to carry the reaction to saturation.[19]

Another interesting phenomenon based on the reaction between clay minerals and amines was reported a few years later.[13] It was found that if clay minerals were mixed with aromatic amines, either in dry powdered

form or in watery suspension, some very pronounced colors were formed. The reaction might be explained as follows. After the amine is brought into close contact with the clay particle, it is adsorbed by the available unsaturated metal groups on the surface. One of the unshared electrons in the nitrogen outer shell is transferred to the crystal structure, setting up an unbalanced force field in the amine. This condition provokes resonance in the amine and a quinoidal structure is obtained. In the case of compounds which exhibit different colors in acidic and basic media, the extent of the resonance will be the controlling factor in color. Benzidine in basic media gives a deep blue color, in acid a yellow color, and in highly acid solutions no color at all. Preliminary tests indicated that the color is determined by the amount of acid salt formed with the amino groups. Thus, in basic media no salt formation occurs and the resonance goes through the two rings of the biphenyl nucleus, resulting in a deep blue. If sufficient acid is added to form salt with one-half the amine present, the resonance is restricted to one ring and, since association is prevented spatially, the yellow color typical of imine compounds is produced. More acid forms salt with all of the amino groups, no resonance is induced, and a colorless state results.

It was later found that it is possible to improve the waterproofness, the humidity resistance, and electrical properties of bodies composed of colloidal crystalline inorganic hydrous oxides, these compounds contain structural water, are capable of swelling when brought into contact with water to form plastic hydrogels, and exhibit base-exchange properties, and their individual crystallites exhibit at least one surface plane in the form of a silica sheet having hexagonal voids with an inscribed circle of about 2.6 Å diameter.[20] This change can be accomplished by contacting the above body with a solution of a salt, the cation of which is of such size that it cannot enter into the crystal lattice of the hydrous oxide and the anion of which contains a chain of from one to six carbon atoms; the salt being capable of forming a water-insoluble basic salt upon being heated. Suitable cations are those of lead, potassium, barium, and ammonium; suitable anions are those of the fatty acids containing up to six carbon atoms, e.g., formic and acetic. Specific examples of these salts or soaps are lead formate, lead acetate, lead methyl ethyl acetate, lead trimethyl acetate and lead dimethyl ethyl acetate.

A subsequent discovery was the replacement of fatty acid anions with anions capable of polymerization, acids such as acrylic, methyl acrylic, vinyl acetic, vinyl acrylic, and in general, substituted acids of this type. Bentonites which have been base-exchanged by such complex ions are rendered water-insoluble and nonswelling because the readily-hydrated sodium ions have been replaced by less hydratable cations like lead. The

water- and humidity-resistance is raised by causing the anions to poly-
merize, thereby giving the surface a highly hydrophobic property.[21]

The next step in this new field was the production of a molding
powder composed of clay made organophilic by reacting its inorganic
cation, i.e., sodium, with the salt of a polymerizable olefinic carboxylic
acid, e.g., lead acrylate. The sodium bentonite may be treated with a
solution of the lead acrylate so as to substitute a complex lead-acrylate
cation for the sodium ion of the bentonite by base exchange and the
resulting bentonite-lead acrylate molecule polymerized to yield a product
having excellent electrical and mechanical properties which render it
suitable for use as an insulating material. It appears that in addition to
the lead-acrylate ion which is introduced into and chemically combined
with the bentonite by base exchange, additional lead acrylate may be
simply adsorbed and/or absorbed on the bentonite particles, and that in
the polymerization this additional lead acrylate becomes chemically com-
bined by polymerization of its acrylic group with those bonded to the
bentonite. Whether or not free lead acrylate is present in the molding
mixture, it will be seen that the polymerization results in a homogeneous
giant molecule containing the bentonite particles, the lead, and the acry-
late groups, all chemically combined and free of interfacial polarization.

A bonding of polymerizable olefinic carboxylic acid compound to the
ion-exchange solid signifies a chemical bonding which may be different
in kind or in degree from that existing between, e.g., sodium and chlorine
in sodium chloride, but which, on the other hand, is something more
than the mere mechanical bond without ion exchange, between the con-
stituents of a simple mixture such as heterogeneous systems of organic
plastics and fillers, e.g., mixtures of asbestos or cellulose fibers, mica or
quartz particles, or the like, with "Bakelite," acrylic resins, or the like.

The differences resulting from such chemical bonding as compared
with simple mechanical mixture can best be illustrated by comparing
data on power factors, obtained with molded test pieces composed of,
e.g., 50 per cent lead clay and 50 per cent lead acrylate, a mechanical
mixture of nonbase-exchanging ground mica (50 per cent) and lead
acrylate 50 per cent, and a mixture of 50 per cent ground quartz and
50 per cent lead acrylate. It will be found that the chemical bonding in
the case of the lead-clay and lead-acrylate mixture eliminates interfacial
polarization, typical for the mechanical mixtures. The absence of inter-
facial polarization is responsible for the low power factor as compared
with those obtained with the other two mixtures.[22]

A systematic study of the products obtained by reacting organic bases,
specifically amines and their salts, with bentonite also revealed that the
resulting product is not only hydrophobic but organophilic and capable
of behaving in organic liquids like the natural clay in water.

Primary aliphatic ammonium bentonite complexes were prepared and gel volumes determined in nitrobenzene, benzene and isoamyl alcohol.[16] Figure 1,[16] in which gel volume is plotted as a function of the number of carbon atoms in the amine chain up to and including eighteen, shows that organophilic properties are negligible until an amine chain of ten carbon atoms is reached and that twelve carbons are required for maximum swelling. Further, it is obvious that nitrobenzene, a liquid of high dielectric constant, is remarkably higher in solvating ability than benzene or isoamyl alcohol.

By applying a Geiger-counter x-ray spectrometer and using 9.6Å as the 001 or basal-plane spacing of the montmorillonite laminae, a stepwise
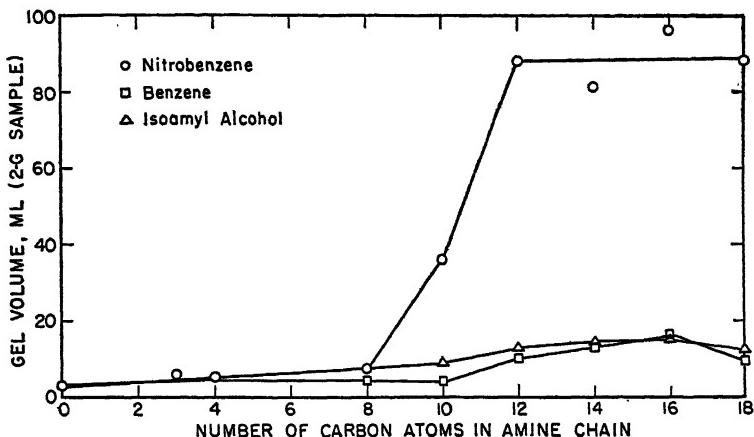
Figure 1. Effect of amine chain length on swelling of organic ammonium bentonites in organic liquids.

separation of the flakes was observed as the length of the amine chain attached to the clay was increased. These steps were of the order of 4Å units or approximately the van der Waals' diameter of a methyl group. Gieseking[3] and Hendricks,[14] using a similar technique, have pointed out that organic molecules attached to montmorillonite through base-exchange reaction tend to be attracted more or less in their entirety onto the surface of the mineral plates, the noncationic portions being held by adsorptive forces. Utilizing Hendricks'[15] demonstration that about 80 per cent of the exchange positions of montmorillonite are on the basal plane surfaces, with the remainder on the edges of the flakes, it was computed that the average area per base-exchange position on the basal plane surfaces should be of the order of 165Å^2 for a clay having a base-exchange capacity of 100 me/100 g. For the relatively simple case of a normal primary aliphatic amine reacted with bentonite, the stepwise separation of the plates by units of 4Å has been taken to indicate that

the hydrocarbon chains lie flat along the surface, with the planes of the zig-zag chains parallel with the plane of the mineral; and the areas covered by such molecules have been calculated and are expressed graphically in Figure 2.[16]

Since the dodecyl complex gave optimum results in three liquids of rather diverse types, experiments were carried out with a wide variety of liquids.[16] The results suggest that solvation may be low in liquids of nonpolar nature, such as the aliphatic and aromatic hydrocarbons. Generally the gel volume appears to increase with the dielectric constant of the liquid, although the correlation is not perfect. A qualitative observa-
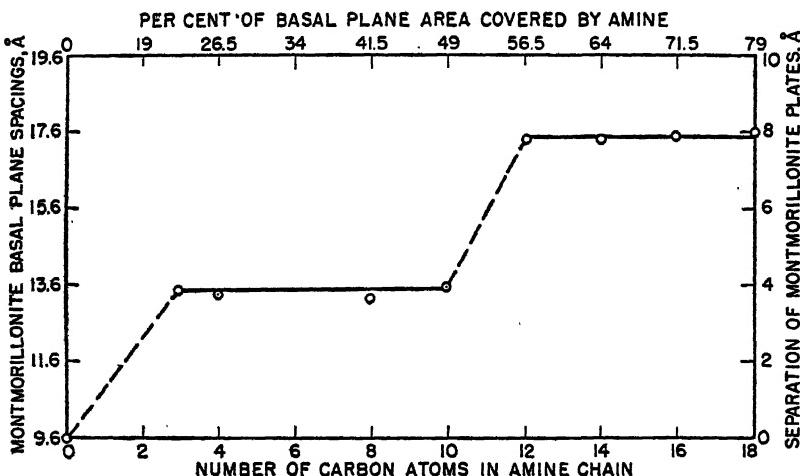
Figure 2. Effect of amine chain length on montmorillonite basal plane spacings.

tion would indicate that the most effective liquids are those which combine highly polar with highly organophilic characteristics, good examples being nitrobenzene and benzonitrile. Acetonitrile, relatively ineffective, satisfies the polarity requirement but is deficient in organic character; on the other hand, toluene, also ineffective, is highly organic in nature but not sufficiently polar.

Since the mixture of toluene with 10 volume per cent of methanol was found to be highly effective in solvating octadecylammonium bentonite, it was used to check the results plotted in Figure 1, regarding the requirements in molecular size of the cation for converting bentonite to the organophilic condition. The striking similarity of this curve (Figure 3,[16] relating gel volume of amine bentonites to amine chain length) to the curve obtained with nitrobenzene confirms the previous finding that solvation is negligible until a chain length of ten carbons is reached or until approximately half of the clay surface is coated.

All the systems tested so far have involved the bentonite salts of single-chain primary amines of less than sufficient size to coat the mineral platelets completely with a single layer of hydrocarbon chains. Bentonite was reacted with quaternary ammonium salts having two long aliphatic chains, and gel volumes were found to be generally good in a single hydrocarbon liquid system. This result bears out the indications that solvation of the complex is a function of its compatibility with the molecules of its liquid environment, and suggests that solvation should take place to a greater extent as the surface loses its inorganic character and becomes more and more like the liquid medium.
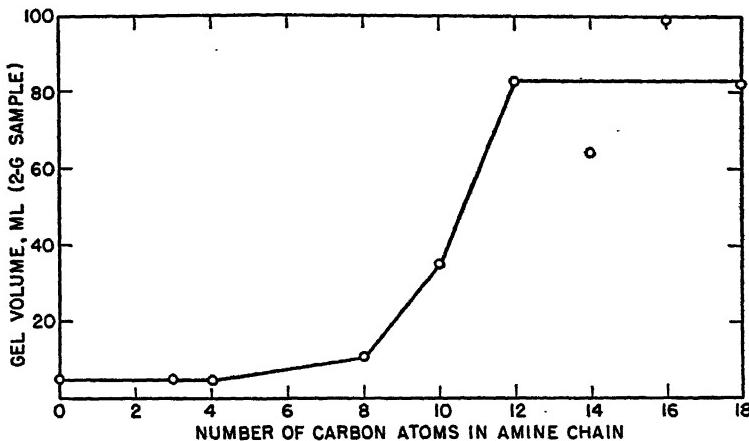
Figure 3. Swelling of organic ammonium bentonites in toluene (90) methanol (10) mixture.

Considering the apparent relationship of surface coverage to the property of swelling in organic liquids, a series of dodecyl ammonium bentonites were prepared in which the ratio of amine added to bentonite was varied over wide limits. Similar series were prepared by the use of octadecyl ammonium acetate and dimethyl dodecyl hexadecyl ammonium bromide. Gel volumes were determined in nitrobenzene as illustrated in Figure 4.[16] From the general similarity of the curves and the coincidence of the gel volume peaks at 100 me/100 g, it is concluded that satisfaction of the base-exchange capacity of the clay is an important factor, since on the basis of surface coverage alone the dialkyl quaternary compound and the octadecyl ammonium compound should achieve optimum results at lower molecular ratios than for the dodecyl compound. These facts lead to the conclusion that the degree of solvation depends upon at least three factors: (1) the extent of the surface coating of the clay particles by organic matter; (2) the degree of saturation of the base-exchange capacity of the clay by organic cations; and (3) the nature of the solvating liquid.

A phenomenon exhibited by these organophilic clays, which deserves special attention both from a scientific and an industrial point of view, is thixotropy. So far only one organophilic thixotropic system has been reported.[18] Mercaptobenzothiazol, if dispersed in benzene, toluene, nitrobenzene, or carbontetrachloride, will yield thixotropic gels at concentrations not below ten per cent. Although this is high in comparison to the one per cent concentration needed for noticeable thixotropy with organophilic bentonite, it must be considered that the bentonite particles are much smaller and therefore offer a far greater surface area per unit volume to the dispersion medium. Even the most recent textbooks on

o Dodecylammonium Bentonite in Nitrobenzene
□ Octadecylammonium   ▪   ▪   ▪
△ Dimethyldodecylhexadecylammonium Bentonite in Nitrobenzene
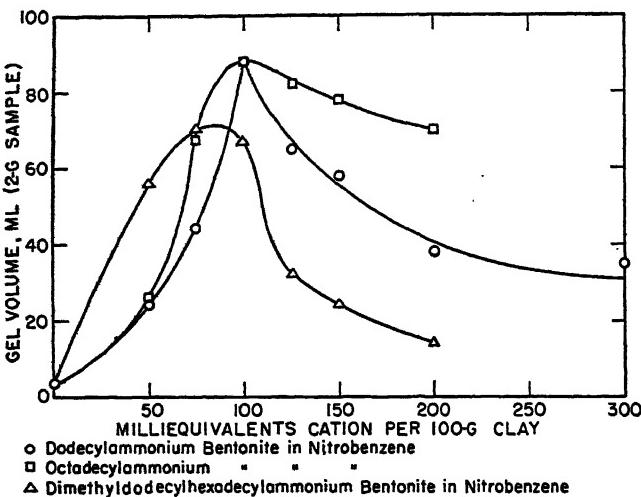
Figure 4. Effect of varying ration of cation to bentonite upon gel volume in organic liquids.

colloid chemistry explain the phenomenon merely by stating that a thixotropic condition exists for certain lyophilic systems when the electrokinetic potential is adjusted at some rather definite position between that which renders the system stable and that which causes coagulation,[4] without even giving an indication that thixotropy might also be the result of the interaction of other than electric forces. This is the more astounding since the only book devoted exclusively to this colloidal phenomenon contains the following statement: "The discussion has been blemished hitherto by a simplification which is really not justified. There exist most likely only a few borderline-cases, where an equilibrium between electrical forces and those of van der Waals is exclusively instrumental in the thixotropic behavior of a gel. In by far the most cases another factor is highly influential which has been neglected so far: it is the influence of hydration, or more generally of solvation. That solva-

tion is of paramount importance is shown most forcibly by the fact that thixotropic pastes have been made from a fine powder (diameter 0.5 to 5μ) of mercaptobenzothiazol and organic insulating liquids such as benzene, toluene, gasolene, etc. Any influence of electrical properties has to be rejected here. Thixotropic behavior is most likely due to a marked affinity between the particles and the liquid, causing the formation of rather thick layers of liquid around the particles. The latter swell, perhaps throughout or to a certain extent; the swollen particles interlock, while the amount of liquid between them is sufficient to allow a thixotropic behavior, when the orientation of the particles is disturbed by shaking." [2]

Another approach towards converting the hydrophilic clay into a clay exhibiting organophilic properties [28] also makes use of the colloid chemistry, structure and composition of these colloidal substances.

Structural analysis of bentonite clays has demonstrated the presence of sheets of silicon and oxygen hexagons which are bonded together by a hydrated magnesium-aluminum layer. It seems logical that the negative charge of the silicate nucleus is concentrated on or near an oxygen atom, since oxygen has chemically a much greater affinity for electrons than has silicon, magnesium, or aluminum. The chief cause for this negative charge is the substitution of magnesium for aluminum. Hence there is good reason for assuming that the negative charge is concentrated on the oxygen.

When the naturally-occurring sodium bentonite is dried, the sodium ions can penetrate through the voids in the outer silicon oxygen sheet and directly satisfy the negative charges. However, if the sodium is replaced by base-exchange reaction with a potassium salt, the situation is changed. Potassium ions having an apparent diameter of 2.66Å, even when dry, cannot penetrate the voids in the silicon-oxygen sheet of the clay. Hence there is considerable distance between the positive charge on the potassium ion and the negative charge in the crystal lattice. It seems as if there are two possible explanations of the situation which actually exists. Either the potassium ion is held weakly in position by the distant charge, or it is possible that the presence of the potassium ion partially or wholly distorts the electronic configuration of the clay so that the charge is pulled to the surface of the crystal.

The condition which exists in a potassium bentonite may be analogous to an organic potassium salt. In potassium propionate, for example, the potassium is joined to an oxygen which is joined to a carbon. Similarly, in potassium bentonite, we have a potassium bonded to an oxygen which is bonded to a silicon. Knowing that the salts of the potassium propionate type will react with such esterifying agents as ethyl iodide to

produce a class of organic compounds known as esters, clay ester could be produced by similar esterification of the potassium bentonite.

The possibility of producing such a material as an allyl clay adds a new method for the production of polymerizable clay or clay particles which may be copolymerized with other monomers such as styrene.

The general equation for the esterification reaction is:

$$RX + \text{clay } M = MX + \text{clay } R$$
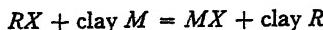
where:

R is an organic radical
X is a halogen
M is a metallic ion

The product of this reaction has been arbitrarily classified as a clay ester. This may not be a strictly accurate definition, since for the purpose of this reaction there is some question as to whether the metallic group on the unreacted clay is attached in a manner similar to that in which a metallic radical is joined to the salt of an acid, or to that of the metallic alcoholate. The former type of bond would upon reaction yield an ester, while the latter would yield an ether. Since the acid-salt bond is generally more of an ionic bond than the alcoholate bond, and since the clay-metal bond is capable of being ionic in water medium, it would seem that the product of reaction of an organic halide with a clay containing replaceable metallic groups would contain a bond that would more closely resemble the ester than the ether linkage. The product was therefore defined as a clay ester.[23]

Applications

The development of organophilic clays by base-exchange reactions, or by esterification, can very well be considered as a new bridge between inorganic and organic matter. It also offers further proof that colloid chemistry is primarily the chemistry of matter characterized by a high surface-over-volume ratio and that the reactivity of substances present in this state is controlled by the composition of its surface.

Organophilic clays can be used for all purposes where a great affinity to an organic dispersion medium is of importance. Paints, varnishes, lubricants and greases exhibiting thixotropic properties can be produced by incorporating organophilic clays into the organic dispersion medium. In combination with inflammable liquids, organophilic clays can cause these products to gel, thus facilitating transportation and storage. Their affinity to organic matter facilitates their incorporation into plastics or rubber. Besides this, the organophilic clay, if properly selected, can be made to combine chemically with the high molecular matrix. If this is accomplished, it becomes a part of the final structure of the polymer and

not just an inert filler. Just as standard bentonite has been of great value in studies of flow in aqueous media,[6, 7] so organophilic clays can now be applied in similar studies which involve the use of organic liquids. These examples should suffice to indicate the tremendous potentialities of this new development in science and technology.

REFERENCES

1. Bart, R., and Gusman, S., "A Study of the Esterification of Bentonite Clays," S.B. thesis, Chem. Eng. Dept., Cambridge, Mass., Mass. Inst. Tech., 1946.
2. Freundlich, H., "Thixotropy," Paris, Herman & Cie., 1935.
3. Gieseking, J. E., *Soil Sci.*, 47, 1 (1939).
4. Hartman, R. J., "Colloid Chemistry," 2nd ed., p. 419, Boston, Houghton Mifflin Co., 1947.
5. Hauser, E. A., "Colloidal Phenomena," New York, McGraw-Hill Book Co., Inc., 1939.
6. ————, "Colloid Chemistry of Clays," *Chem. Revs.*, 37, 287 (1945).
7. ————, and le Beau, D. S., "Colloid Chemistry of Clay Minerals and Clay Films," in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. VI, New York, Reinhold Publishing Corp., 1946.
8. ————, and ————, "Silicic Chemistry," New York, D. Van Nostrand Co., Inc., 1951.
9. ————, and Johnson, A. L., *J. Am. Ceramic Soc.*, 25, 223 (1942).
10. ————, and Reed, C. E., *J. Phys. Chem.*, 40, 1169 (1936).
11. ————, and ————, *J. Phys. Chem.*, 41, 911 (1937).
12. ————, and ————, *J. Am. Chem. Soc.*, 58, 1822 (1936).
13. ————, and Legett, M. B., *J. Am. Chem. Soc.*, 62, 1811 (1940).
14. Hendricks, S. B., *J. Phys. Chem.*, 45, 65 (1941).
15. ————, Nelson, R. A., and Alexander, L. T., *J. Am. Chem. Soc.*, 62, 1457 (1940).
16. Jordan, J. W., *J. Phys. Coll. Chem.*, 53, 294 (1949).
17. Marshall, C. E., and Krinbill, C. A., *J. Phys. Chem.* 46, 1077 (1942).
18. Recklinghausen, H. v., *Koll.-Z.*, 60, 34 (1932).
19. Smith, C. R., *J. Am. Chem. Soc.*, 56, 1561 (1934).
20. Hauser, E. A., U. S. Patent 2,317,685 (Apr. 27, 1943).
21. ————, U. S. Patent 2,383,647 (Aug. 28, 1945).
22. ————, and Dannenberg, E. M., U. S. Patent 2,401,348 (June 4, 1946).
23. Bart, R., and Gusman, S., Private communication.
24. Way, J. R., *J. Roy. Agr. Soc., Engl.*, 11, 13 (1850).

THEORY OF COLLOIDAL BEHAVIOR OF SOILS

Amar Nath Puri *

Field Research Station, Bombay, India

Base-Exchange Phenomenon

Adsorption of bases by soils was first observed almost simultaneously by Way[1] and Thompson[2] around 1850. Both these authors ascribed the phenomenon to simple chemical action. Liebig,[3] in 1855, first criticized and later[4] confirmed Way's data. He believed that a physical force held the bases in capillaries, more or less as water is held in a sponge. Eichhorn,[5] in 1858, working with chabazite, contradicted Way's results regarding the order of the replacing power of calcium and sodium. Later on, Henneberg and Stohmann[6] confirmed some of Way's conclusions. Peters[7] showed a relationship between adsorption of bases and concentration of the solution and concluded that the former was due to surface attraction and was therefore essentially physical in nature. Frank[8] noted that acid soils imparted an acid reaction to salt solutions. This observation, later confirmed by others, led him to believe that base-adsorption was purely a physical matter—i.e., the cation of the salt was adsorbed, leaving the anion in solution.

Heiden[9] observed that the quantity of magnesium adsorbed from $MgSO_4$ was approximately equivalent to the sum of the calcium, potassium, and sodium replaced, and he therefore concluded that the adsorption phenomenon was mainly chemical. Knop[10] correlated the NH_4-adsorbing power of soil with the amount of fine-grained material present, which was also Way's conclusion. Beidermann[11] found this correlation good with some soils, but not with others. Pillitz,[12] apparently for the first time, introduced the concept of maximum adsorptive power of soils leached with different concentrations of NH_4Cl under constant conditions of temperature. He maintained that the function of the added base was to neutralize the soil acids.

In 1876, Lemberg [13] showed that crystalline silicates other than zeolites also exchange cations with salt solutions. More recently, Vanselow confirmed Lemberg's results.[14] In 1878 and 1879, Van Bemmelen published two papers [15] in which he showed that the exchange of bases in soils is a truly chemical reaction. Thus, he not only joined the chemical school of thought, but supported this contention more vigorously than any of the previous workers had done. In 1888, however, he revised his views radically [16] and advocated the colloidal concept of adsorptive power of soils. He recognized the presence of colloidal silicates—iron oxide, silicic acid, and organic substances—and considered these hydrogels chiefly responsible for adsorption; however, the reaction between zeolites and salt solutions was still thought of as chemical by him.

Following the publication of Van Bemmelen's papers, there was little progress in soil science along theoretical lines until in 1912 interest was again revived by Wiegner.[17] It is difficult to obtain a clear picture of the theory of soils from the earlier workers; certainly Wiegner's many papers on the subject did not indicate whether he believed in physical adsorption or ordinary chemical reaction. The next twelve years, from 1912 to 1925, were marked by the outstanding researches of Gedroiz. His numerous papers, originally published in Russian, were translated by Waksman, and mimeographed and distributed by the United States Department of Agriculture. Gedroiz recognized the importance of the specific surface of the solid phase in regulating the energy of adsorption and found it difficult to decide whether adsorption (or exchange of bases) was purely a physical or chemical reaction, and he therefore favored a physicochemical interpretation. His main contention was that "if the solid phase is sufficiently finely divided, it is not necessary at all for it to go into solution in order for the reaction with a substance in solution to take place energetically. But such reactions can take place only on the surface of the solid."

Mattson, in a series of papers published in *Soil Science* under the general title, "Laws of Soil Colloidal Behavior," has tried to explain the base-exchange phenomenon in terms of acidoid-basoid ratio, i.e., the molecular proportions of silicic acid and hydroxides of aluminum and iron. In the course of his researches extending over a period of about twenty years, he found it necessary to alter his views in several respects. He discussed various aspects of soil colloidal behavior such as base exchange, Donnan equilibrium, isoelectric point, coagulation, and soil formation. However, it is not possible to build up a comprehensive and coherent theory of the behavior of soil colloids from his papers. Mattson's main thesis was that the inorganic colloidal materials of soil were of the nature of isoelectric precipitates; not mere mechanical mixtures of silicic acid and hydroxides of aluminum and iron, but chemical compounds

resulting from the chemical union of one or more of the acidoid H ions and basoid OH ions and the splitting off of water. The acidic behavior of the soils was explained on the basis of the residual H ions of the silicic acid. Bradfield [18] applied the acidoid concept of Michaelis to soils. This view was further elaborated by Hissink,[19] Page,[20] and Puri.[21] They compared soil colloidal material to insoluble weak acids. Hence the soil could be fully base-saturated only at a high pH value.

Comparatively recently, the x-ray analyses of various minerals found in soils have been cited as evidence that the exchangeable cations are attracted to the crystal lattice by its excess of negative charges. Kelley,[22] Hofmann et al.,[23] Marshall,[24] Ross and Hendricks,[25] and others have extended this knowledge, particularly regarding the crystal structure of clay in relation to base exchange. The importance of the crystalline structure of soil minerals has caught the imagination of soil chemists to the extent that the colloidal state is receding into the background, and greater and greater emphasis is being laid on the nature of the exposed surfaces.

Mechanical and Ultramechanical Analyses of Soils

Side by side with the development of these physicochemical theories, the mechanical analysis of soils was gaining popularity. The term, mechanical analysis, was first introduced in 1805, although the principles and early methods for this type of analysis were known in ancient Greece, and sieves were used to separate sands in 1704. However, not until the discovery of Stokes' law in 1851 and its application to the mechanical analysis of soils in 1867, was any importance attached to these determinations.

The concept of mechanical analysis and its importance in soil studies received great impetus from the work of Odén [26] who, in collaboration with Fisher, perfected the theory of sedimentation analysis.[27] This method, introduced at the same time as the pipette method of mechanical analysis,[28] became a standard, routine procedure for soils. Several other procedures and apparatus were devised, but it was soon realized that most of the newer methods rested on the assumption that the soil suspension prior to analysis represented primary particles. This was far from correct, for soil aggregates were found to be notoriously resistant to dispersion methods, and in spite of the cooperative work carried out under the auspices of the International Society of Soil Science, the problem remained unsolved and no single method could be recommended to disperse completely all types of soils. Thus, even at present, various laboratories have adopted different methods of dispersion to suit their individual requirements.

There was another aspect of mechanical analysis around which a good deal of controversy raged in soil literature, namely the limiting size of the various fractions obtained on analysis. This was particularly important in fixing the exact size of what is known as the clay fraction. The average diameter of the clay fraction has been arbitrarily fixed at 0.005 mm (American), 0.001 mm (English), and 0.002 mm (Atterberg). The Atterberg size has been recommended for international adoption. It is evident that even the smallest diameter of clay particles is much too coarse as compared with the size attributed to colloidal particles. Besides, since the specific surface (surface area per unit mass of the particles) is given by the formula:

$$SS = 0.02264 \Sigma p/D$$

where p is the percentage by weight of particles of mean diameter D, the magnitude of this value is profoundly affected by the size of the lower limit of analysis. For instance, a simple calculation will show that if the mechanical analysis of a soil is stopped at 0.001-mm diameter, for particles which might easily represent aggregates of particles of 0.00001-mm diameter, our calculated value of specific surface would be only one per cent of the true value.

It is inconceivable that natural forces of weathering should have abruptly stopped at 0.001-mm particles; therefore the existence of particles of smaller dimensions, or ultraclay, is only a logical deduction from the nature of the size-distribution curves of soils. Curiously enough, no serious attempt was made to determine the size-distribution of soils in the region of ultraclay until Marshall [28] devised a centrifugal technique, in which a thin layer of the soil suspension from which all the coarser particles had been removed was put on top of a thick layer of a denser liquid, such as sugar or urea solution. Under these conditions, according to Marshall, the soil particles reaching the bottom of the tube between any two given periods of time all lay within the two limits of diameter calculated from Stokes' law. Marshall, however, did not report the ultra-mechanical analysis of an appreciable number of soils.

The simplicity of gravity sedimentation and the application of the pipette method to the ultramechanical analysis of soils, although pointed out by Robinson as early as 1922,[29] was never seriously considered until 1941, when Puri and Puri [30] devised a micropipette technique which they compared with that of Marshall, as well as with the ordinary pipette method, and showed general agreement. When dealing with ultraclay particles it is more convenient to plot the negative indices of the diameters, in centimeters, against the summation percentages. These values are conveniently representd by the symbol pD. The advantage of this mode of representation has been discussed by Puri.[31]

One of the greatest obstacles in the way of the development of a comprehensive theory of soils has been the lack of knowledge of ultramechanical analysis. Since there was no method of verifying the extent of the total surface presented by the soil, all correlations of adsorptive capacity with clay content were of empirical nature and were disproved or discarded in the course of time. Thus, the controversy as to whether adsorption of bases by soils is a chemical or physical phenomenon still goes on, and no comprehensive theory of the colloidal behavior of soils has resulted from these discussions.

Moisture Adsorption of Soils

Agricultural chemists were concerned with another aspect of the soil besides mechanical analysis and physicochemical properties, namely, the adsorption of moisture. The majority of investigators of soil moisture concerned themselves with its distribution and movement in the liquid state; its adsorption from the vapor phase received very little attention until a much later stage.[32-34] Whereas in studies of movement and distribution of water the total interstitial space is operative, the vapor pressure of soils at different moisture contents is very largely controlled by the minute pores associated with ultraclay; the larger voids have comparatively little influence.

Since water is one of the most important factors for the growth of crops, attempts were made in the early days to find easily determined equilibrium points that would define the power of the soil to hold water. Thus such empirical determinations as normal moisture capacity, field capacity, sticky point, liquid limit, plastic limit, minimum water of saturation, moisture equivalent, wilting coefficient, and hygroscopic coefficient all enjoyed their due share of popularity.

In recent years there has been a tendency to abandon the so-called divisions of water. It is tacitly assumed that all water is held by the same capillary force, the magnitude of which depends on the size of the pores. Thus Schofield has expressed the moisture content of soils in terms of free-energy relationship.[35] If we suppose that water is held by soil against a suction force tending to displace it, the free energy can be expressed in terms of the height, in centimeters, of the equivalent water column. In order to deal conveniently with the whole range of suction, use is made of the logarithm of this height. By analogy with Sorensen's logarithmic acidity scale, the symbol pF has been used to express the magnitude of the suction force (F being the recognized symbol for free energy). Following this line of thought, Puri has shown that pF is numerically equal to pD, which as mentioned before is the negative index of the diameter of particles, in centimeters.[31] The extension of this rela-

tionship in the region of ultraclay has been made possible by the following formula given by Schofield:

$$pF = 6.5 + \log_{10}(2 - \log_{10} h)$$

where h is the relative humidity with which the soil is in equilibrium. A simple calculation with the help of this formula, coupled with the fact that pF is numerically equal to pD, will show that particles of 0.001-mm diameter have interstitial capillaries which could only adsorb water from an atmosphere of relative humidity not less than 99.5 per cent. Thus, conventional clay which marks the lower limit of ordinary mechanical analysis would cease to interest us as regards moisture adsorption. However, the existence of particles smaller than 0.001 mm is proved not only by the fact that soils can adsorb moisture from humidities as low as 10 per cent, but as pointed out before, the presence of these particles can be demonstrated by extending the mechanical analysis of soils to the region of ultraclay.

Theory of Soil Colloidal Behavior

Base exchange, mechanical analysis, and moisture adsorption are the three main topics of soil science about which there was much controversial discussion in the literature. With the knowledge of ultramechanical analysis of soils, we can now present the theory of soil colloidal behavior which links up the above three major phases of soil science and offers a rational explanation for all the complex phenomena associated with soils.

The basic constituents of inorganic soils are ferroaluminosilicates which, if molecularly dispersed, would behave like any soluble acid. The acidic hydrogen of the primary silicic acid combines but feebly with the hydroxides of iron and aluminum, so that qualitatively the product is no different from pure silicic acid. If anything, its hydrogen-ion activity is slightly enhanced thereby. Thus, if we take a soil from anywhere in the world and remove from it all the extraneous materials, like salts, by leaching with dilute hydrochloric acid, it will behave in every way like an acid. No matter what the nature of the original soil, every soil so treated behaves like any other soil qualitatively, namely, it is a water-insoluble acid like other water-insoluble acids such as stearic, uric, or myristic acids.

An acid, according to the classic definition, is a substance which contains hydrogen replaceable by a metal. This hydrogen is not adsorbed or held in any other way, but constitutes an integral part of its molecular make-up. The definition of an acid makes no distinction between soluble and insoluble acids. As is well known, a number of acids like stearic, uric,

or silicic acid are insoluble in water. However, these insoluble acids can be dispersed in water to such an extent that they can be titrated with an alkali exactly like any water-soluble acid.[36] The distinction between a soluble and an insoluble acid, therefore, is purely arbitrary and the acidic hydrogen, whether in molecular state or constituting the surface of a solid, behaves exactly the same.

If the molecules constituting the surface of a solid are designated as its "active mass," every substance in contact with water would tend to behave chemically like a true solution to the extent of its active mass. For instance, if one gram of a substance containing 10 per cent active mass is suspended in 100 cc of water, it will be equivalent to a true solution of 0.1 per cent concentration. Every solid is "dead" chemically. As it is broken down to smaller mits, or rolled into thin plates, or drawn into a wire, so that fresh surfaces are exposed, its active mass increases and ultimately becomes 100 per cent when the solid is reduced to molecules, monomolecular films, or filaments. The important consideration is that the "exposed" molecules of every substance are "live" and active chemically and physicochemically, and enter into chemical reactions characteristic of those molecules exactly as if they were in solution.

As pointed out elsewhere, specific surface of soils can be determined from ultramechanical analyses, assuming a spherical shape for the particles and an "equivalent diameter" in accordance with Stokes' law.[37] Specific surface can also be determined from the vapor-pressure curve of the soil, by means of the formula previously noted.[38]

$$pD = pF \times 6.5 + \log_{10}(2 - \log_{10} h)$$

Thus, we have two independent methods of finding the specific surface of soils. If the average diameter of ferroaluminosilicates is assumed to be equal to 8Å, we can calculate from the specific surface of the soil its active mass, assuming a uniform distribution of the molecules on its surface. This calculated value can be compared with the base equivalent of the soil, determined from its titration curve.[39] In Table 1 a comparison of these values for a number of soils is given by way of illustration. These values have been selected at random from a larger collection which included soils of practically all types.

Objections may be raised to the assumption that the particles are spherical. A uniform diameter of the ferroaluminosilicates is also problematical, nor can the same equivalent weight be assumed for every unit active mass of silicates of diverse origin. Nevertheless, the values are sufficiently close to justify the belief that natural forces of weathering have produced, on the average, somewhat of a similar effect during the disintegration of parent rocks. A better correlation may be expected

within a group of soils of the same mineralogical composition. When all soils are lumped together, a better agreement cannot be expected.

The following additional evidence, based on the examination of a number of soils, further supports the contention that the colloidal chemistry of soils is the chemistry of acids and bases, that the soil-water system is in effect a "soil solution," and that every soil suspension behaves like a true solution of an acid or its salt to the extent of its active mass or exposed molecules constituting the total surface of the particles.

TABLE 1. SPECIFIC SURFACE OF SOILS, OBTAINED FROM
DETERMINED (S) AND CALCULATED (S') RESULTS OF
ULTRAMECHANICAL ANALYSIS, AND FROM OTHER DATA

Soil No. P.C.	Specific Surface (sq m/gm)		Base Equivalent (m. e./gm)	
	S	S'	B	B'
2	151	238	97.8	108.8
3	194	233	106.8	122.0
4	16	28	11.0	12.6
5	16	18	8.5	21.1
6	67	49	29.0	23.6
8	102	72	43.5	39.2
9	38	37	18.8	15.4
10	110	98	52.0	39.6
11	111	108	54.8	52.0
12	22	12	8.5	9.8
13	180	206	96.5	80.0
14	81	40	30.2	48.0
16	7	9	4.0	11.4
17	30	15	11.2	13.0
18	20	34	13.5	19.8

S = specific surface obtained from determined ultramechanical analysis
S' = specific surface calculated from the vapor pressure curve
B = base equivalent calculated from the mean of S and S'
B' = base equivalent determined by titration

(1) Every soil, to the extent of its active mass, can be titrated with any alkali, leading to a well-defined titration curve which can be reproduced backward and forward at will.[39-41]

(2) It can decompose carbonates with the evolution of CO_2, the decomposition proceeding quantitatively in stoichiometric proportion to the active mass.[42]

(3) It can decompose sulfides with the evolution of H_2S quantitatively and stoichiometrically (Ref. 38, *loc. cit.*, and Part I, Chap. 5).

(4) It can hydrolyze esters and bring about the inversion of sucrose strictly in accordance with the catalytic activity of its hydrogen ions.[43]

(5) Heat of neutralization of soil acidoids follows the same pattern as that of ordinary soluble acids (Ref. 38, *loc. cit.*, and Part I, Chap. 8).

(6) Action of neutral or hydrolyzed salts can be interpreted fairly accurately on the basis of distribution of a base between two acids

of unequal strength, in which the active mass of the insoluble acidoid behaves exactly like a soluble acid.[44]

(7) Decomposition of soils by electric current into acidoid and hydroxide of the cation proceeds exactly like the electrolysis of common soluble salts.[45, 46]

(8) Neutralization of soil acidoid by two different bases, one after the other, leads to the same equilibrium condition, whichever base is added first.[47]

(9) When neutralized with alkalies, every soil acidoid gives stable saloids which are similar for the same saloid of all soils, without exception. For instance, sodium saloids of all soils are completely dispersed at pH 11 on coming in contact with water, without any shaking. On the other hand, calcium saloids of all soils cannot be dispersed in water unless violent disruptive forces are used to break the compound particles into primary units; even then the dispersion does not extend beyond particles of 0.001-mm diameter.[48]

(10) Calcium saloids of soil acidoids in water suspension can be precipitated quantitatively with a carbonate,[49] or an oxalate,[50] strictly in accordance with the standard procedure for the precipitation of soluble calcium salts in analytical chemistry.

(11) Magnesium saloids of soil acidoids in water suspension are quantitatively precipitated as phosphate only in the presence of ammonia, as in the standard procedure of analytical chemistry for the estimation of soluble magnesium as pyrophosphate (Ref. 38, *loc. cit.*, and Part I, Chap. 21).

These facts lead us to the conclusion that all natural soils are mixtures of saloids of soil acidoid, representing single points on their titration curves, which determine the stage to which they are neutralized. Since they are the saloids of a weak acidoid and comparatively strong bases, they are easily hydrolyzed on coming in contact with water. The extent of this hydrolysis must obviously depend on the state of neutralization, which determines the prevailing reaction. Therefore, all natural soils tend to come to a state of equilibrium governed by the quantity of rainfall in that region, and the extent of leaching that follows thereby. Thus, the higher the rainfall in any locality, the lower the prevailing pH value of the soils of the place. Acid soils are, therefore, characteristic of humid regions, and alkali soils, of areas devoid of rainfall.

Plant Growth and Nutrition in Soils

Since all chemical reactions characteristic of soluble acids can be reproduced quantitatively with every soil, which at any time represents an

acid-base equilibrium point on its titration curve, and since the soil-water system is in effect a homogeneous two-dimensional soil solution, plants must grow in soils as they do in nutrient solutions. The mechanism of food intake by the roots is probably even more efficient for the adsorption of nutrient materials from the soil than from the salts in solution. Note, for instance, the observation of Stout and co-workers on the use of radioactive tracers in plant nutrition studies.[51] They found that when the inert cations in the adsorbed state in suspensions of clay were placed in contact with the roots, much greater releases of phosphorus [42] were noted than for the ordinary salt. Again, Nelson and co-workers,[52] using radioactive tracer technique, have shown that native phosphates in soils are better utilized by crops than the added soluble phosphates. They concluded that it is only in the absence of native phosphates that plants show the highest intake of phosphates applied as fertilizer.

A correct appreciation of the mechanism of plant nutrition is of the highest importance in agriculture and food production. If the soil-water system is in effect a "soil solution" subject to the well-known laws of acid–base equilibrium, then it will not be long before we learn how to utilize the nutrient materials locked up in the soil itself. However, basic concepts must be clarified before progress can be made. Results of the highest practical value can only result when the fundamentals of soil science have been firmly established, and it is towards this objective that the present paper is directed.

REFERENCES

1. Way, J. T., *J. Roy. Agr. Soc. Engl.*, **11**, 313 (1850).
2. Thompson, H. S., *J. Roy, Agr. Soc. Engl.*, **11**, 68 (1850).
3. Liebig, J. von, *Ann. Chem. Pharm.*, **94**, 373 (1855).
4. ———, *Ann. Chem. Pharm.*, **105**, 109 (1858).
5. Eichhorn, H., *Poggendorff Ann. Physik Chemie*, **105**, 126 (1858).
6. Henneberg, W., and Stohmann, F., *Ann. Chem. Pharm.*, **107**, 152 (1858).
7. Peters, C., *Landw. Vers. Sta.*, **2**, 131 (1860).
8. Frank, A., *Landw. Vers. Sta.*, **8**, 45 (1866).
9. Heiden, E., *Landw. Vers. Sta.*, **11**, 69 (1869).
10. Knop, W., "Die Bonitirung der Ackererde," 2nd ed., Leipzig, 1872.
11. Beidermann, R., *Landw. Vers. Sta.*, **15**, 21 (1872).
12. Pillitz, W. H., *Z. anal. Chem.*, **14**, 55 (1875); *ibid.*, 283.
13. Lemberg, Z. *deut, geol., Ges.*, **28**, 519 (1876).
14. Vanselow, A. P., *Soil Sci.*, **33**, 95 (1932).
15. Van Bemmelen, J. M., *Landw. Vers. Sta.*, **21**, 135 (1878); **23**, 265 (1879).
16. ———, *Landw. Vers. Sta.*, **35**, 69 (1888).
17. Wiegner, G., *J. Landw.*, **60**, 11 (1912).
18. Bradfield, R., *J. Am. Chem. Soc.*, **45**, 2669 (1923).
19. Hissink, D. J., *Trans. Faraday Soc.*, **20**, 551 (1924).
20. Page, H. J., *Proc. 2nd Intern. Congr. Soil Sci. (Groningen)*, A, 232 (1926).
21. Puri, A. N., *Mem. Dept. Agr. India, Chem. Ser.*, XI (1), (1930).
22. Kelley, W. P., Dore, W. H., and Brown, S. M., *Soil Sci.*, **31**, 25 (1931).

23. Hofmann, U., Endell, K., and Wilm, D., *Z. Krist.*, **86**, 340 (1933).
24. Marshall, C. E., *Z. Krist.*, **91**, 433 (1935).
25. Ross, C. S., and Hendricks, S. B., *Soil Sci. Soc. Am., Proc.*, **6**, 58 (1941).
26. Odén, S., *Soil Sci.*, **19**, 1 (1925). See Odén, S. in "Colloid Chemistry, Theoretical and Applied," Vol. I, New York, Reinhold Publishing Corp., 1926.
27. Fisher, R. A., and Odén, Sven, *Proc. Roy. Soc. Edinburgh*, **44**, 98 (1923-24).
28. Marshall, C. E., *Proc. Roy. Soc. London*, **126A**, 427 (1930).
29. Robinson, G. W., *J. Agr. Sci.*, **12**, 302 (1922).
30. Puri, A. N., and Puri, B. R., *J. Agr. Sci.*, **31**, 171 (1941).
31. ———, *Soil Sci.*, **48**, 505 (1939).
32. Thomas, M. D., *Soil Sci.*, **11**, 409 (1921); 17, 1 (1924).
33. Puri, A. N., Crowther, E. M., and Keen, B. A., *J. Agr. Sci.*, **15**, 68 (1925).
34. Crowther, E. M., and Puri, A. N., *Proc. Roy. Soc. London*, **106A**, 332 (1924).
35. Schofield, R. K., *Trans. 3rd Intern. Congr. Soil Sci.*, **2**, 37 (1935).
36. Puri, A. N., and Bharihoke, G., *Soil Sci.*, **67**, 331 (1949).
37. ———, Rai, B., and Malhotra, J. K., *Proc. Lahore Phil. Soc.*, **6**, 54 (1944).
38. ———, "Soils: their Physics and Chemistry," Part III, Chap. 5, New York, Reinhold Publishing Corp., 1949.
39. ———, and Asghar, A. G., *Soil Sci.*, **45**, 359 (1938).
40. ———, and ———, *Soil Sci.*, **45**, 477 (1938).
41. ———, and Sarup, A., *Soil Sci.*, **48**, 77 (1939).
42. ———, and Puri, M. L., *Soil Sci.*, **46**, 401 (1938).
43. ———, and Dua, A. N., *Soil Sci.*, **46**, 113 (1938).
44. ———, and Uppal, H. L., *Soil Sci.*, **49**, 155 (1940).
45. ———, and Hoon, R. C., *Soil Sci.*, **43**, 305 (1937).
46. ———, and ———, *Soil Sci.*, **44**, 399 (1937).
47. ———, and ———, *Soil Sci.*, **45**, 309 (1938).
48. ———, and Rai, B., *Indian J. Agr. Sci.*, **14**, 210 (1944).
49. ———, *Soil Sci.*, **42**, 47 (1936).
50. ———, *Soil Sci.*, **40**, 383 (1935).
51. Stout, P. R., *et al.*, *Soil Sci. Soc., Proc.*, **12**, 91 (1948).
52. Nelson, L. B., *et al.*, *Soil Sci. Soc. Am. Proc.*, **12**, 113 (1948).

COLLOID ASPECTS OF OIL-WELL TREATMENT

P. H. Cardwell

Dowell Incorporated, Tulsa, Okla.

VARIOUS COLLOIDAL PHASES of the petroleum industry have been covered in the previous volumes of this series of "Colloid Chemistry." Among the different subjects which have been discussed are: petroleum,[42, 55, 103] petroleum emulsions,[55, 100, 103, 107] refining,[55, 103, 107] lubricating,[12, 103, 107] and drilling fluids.[70, 107] This paper will discuss the colloidal aspects of the so-called "down-hole" chemical treatment of oil and gas wells, that is, the placement of chemicals into the well when it is about to be or has been put on production.

Acidizing

The greatest use of chemicals in the down-hole treatment of oil and gas wells is the addition of hydrochloric acid to the oil- or gas-bearing zones of the wells to increase the permeability of the formation and thereby increase production.[114] This has developed into an important use of hydrochloric acid, in fact, in one three-month period in 1940, a total of 1,538,500 gallons of acid were used for acidizing of the Monroe Gas Field in Northeast Louisiana.[137] Many wells are producing from limestone and dolomite formations [26, 28, 37-39, 81, 122, 126, 131] which are soluble in acid. Other wells are in sand formations, and these can be acidized with a mixture of hydrochloric and hydrofluoric acids.[19, 41, 82, 102]

Before the acid contacts the formation to be acidized, it must be delivered to the well in steel equipment and then displaced down the tubing or occasionally down the casing to the bottom of the well. Since, in these operations, the steel equipment is in contact with the acid solution, it is necessary that an inhibitor be used to protect the steel from acid attack.

Inhibitors. In order to understand the theories as to how inhibitors function, it is necessary to examine the complexities of a steel surface.

A steel surface is composed of cathodic and anodic areas. The cathodes are at the points of the crystal boundaries, the slag inclusions, and the location of other impurities, whereas the anodes are the exposed iron surfaces. When acid attacks steel, iron is dissolved at the anodes as ferrous ions and hydrogen is formed at the cathodes.

The inhibitors function by decreasing the formation of hydrogen at the cathodes, which reduces the amount of iron dissolved by the acid. This is accomplished by one of two means, depending upon which of the two types of inhibitor is being used. The two different kinds of inhibitors employed in the hydrochloric acid for oil-field acidizing are the inorganic and organic types. The inorganic inhibitors, of which arsenic is the best example, protects the steel surface from acid attack by plating onto the cathodic areas, forming a new metallic cathode. This new cathode has a higher overvoltage toward hydrogen than the original surface, which brings about a reduction in the evolution of hydrogen.[31]

The organic inhibitors are usually nitrogen or sulfur compounds such as the amines, the nitrogen-ring compounds, and the nitrogen-sulfur compounds, the thioureas. These materials are adsorbed on the cathodic areas of the metal, which increases the hydrogen overvoltage.[29, 70, 91, 92, 93, 104, 121] This adsorbed film of the inhibitor acts as an interfacial resistance, at the cathodic areas, to the evolution of hydrogen. The latest work reported [11] on inhibitors seems to indicate that the adsorbed film does not give mechanical protection, also that the ohmic overvoltage of the film is negligible, but inhibitors seem to give an increase in the hydrogen-activation overpotential.

Since the inhibitors are adsorbed upon the metal surface from an aqueous solution, it is of interest to study this adsorption in terms of the efficiency of the inhibitor. The usual procedure in adsorption work is to determine the actual amount of a substance adsorbed by a certain amount of a solid. In the case of inhibitors it is possible to study adsorption by means of the reduction in the acid attack upon the metal as brought about by the presence of an inhibitor.

The efficiency of an inhibitor can be measured by the corrosion rate of a metal exposed to an acid solution containing the inhibitor. This can be expressed as the percentage of inhibition by the use of the following equation:

$$\% \text{ Inhibition } = \frac{R_1 - R_2}{R_1} \times 100$$

where R_1 = corrosion rate in acid with no inhibitor
         R_2 = corrosion rate in acid containing a specific inhibitor

The efficiency of an inhibitor depends upon the percentage of the total cathodic areas which it covers with the adsorbed film. The greater

the portion of the cathodic areas covered, the higher the efficiency of the inhibitor. The area covered by an inhibitor molecule is the cross-sectional area of the molecule as projected onto the metal surface. In a homologous series with the same angle of orientation of all the molecules to the metal surface, the amount of the cathodic area covered should be directly proportional to the amount of the inhibitor adsorbed. Thus, it is usually assumed in a homologous series with the same angle of orientation of the molecules, that the amount of an inhibitor adsorbed is proportional to the efficiency of the material.

It has been found,[30] in a study of efficiency of a homologous series of aliphatic amines, that Traube's rule of adsorption holds fairly well. In the case of methyl-, ethyl-, and propylamines the rule holds much better than for the higher-member compounds. The longer-chain compounds seem to be better inhibitors than what normally might be expected from Traube's rule. This has been explained [30] from the standpoint that the longer-chain molecules are inclined to the metal surface more than are the shorter ones.

Since the mechanism by which organic inhibitors function is by adsorption upon the metal surface, it would seem that Langmuir's adsorption-isotherm equation [78] would be applicable. The familiar form of Langmuir's equation is:

$$X = \frac{abC}{1 + aC}$$

where
$$X = \text{amount adsorbed}$$
$$C = \text{concentration}$$
$$a, b = \text{constants}$$

This equation can be rearranged into the following form:

$$\frac{C}{X} = \frac{C}{b} + \frac{1}{ab}$$

The term $\frac{C}{X}$ can be plotted against C, giving a straight line, with the reciprocal of the slope equal to b, the amount of maximum adsorption.

If it is assumed that the amount of adsorption is directly proportional to the percentage of inhibition, Ch'iao and Mann [30] have rewritten the Langmuir equation in the following form:

$$\frac{C}{I} = \frac{C}{b'} + \frac{1}{a'b'}$$

In this equation, I is the percentage of inhibition at the concentration C of the inhibitor, and b' is a constant which is the maximum percentage of inhibition.

It has been found [30] that a straight line is obtained when $\frac{C}{I}$ is plotted against C for a number of aliphatic and aromatic amines. The constant b', the reciprocal of the slope, is in the neighborhood of 100. This would be expected since b' is the maximum percentage of inhibition. A few inhibitors such as dibutylamine and quinoline do not give straight lines at low inhibitor concentrations. Here again it seems that the amount adsorbed is not directly proportional to the percentage of the cathodic areas covered owing to the fact that the cross-sectional projected area of the molecules onto the surface varies with the amount adsorbed. Ch'iao and Mann [30] have suggested that a critical concentration of the inhibitor in the acid solution must be reached before the cross-sectional area of the molecules projected onto the surface becomes constant and directly proportional to the amount adsorbed. Below the critical concentration, the molecules incline to the surface at different angles and the effective molecular area is not constant, but decreases with increasing concentration of inhibitor. Langmuir's adsorption-isotherm equation can be applied to percentage-inhibition data only when the amount adsorbed is directly proportional to the percentage of the cathodic areas covered.

Recently it has been reported [20] that wetting agents bring about a reduction in the rate of corrosion of steel in hydrochloric-acid solutions inhibited with various nitrogen bases and thiourea compounds. The wetting agents used were of the anionic type and in themselves were not corrosion inhibitors, but the combinations of wetting agent and organic inhibitors were considerably better than the organic inhibitors alone. Table 1 contains data as reported by Cardwell and Eilers [20] on the effect of wetting agents.

The contact angles and the surface tensions of the 2.5 per cent hydrochloric-acid solutions containing the wetting agent and the various organic inhibitors are not reported in the table. The contact angles when the wetting agent was present were always zero regardless of the organic inhibitor, and the surface tensions of these solutions varied from 29.3 to 29.7 dynes per cm. While it is true that the corrosion rates were determined at 79.4°C, the surface tensions were measured at the acid-air interface at 25°C, and the contact angles were measured at the steel-acid-hydrogen interface at 25°C; it is believed that the information can be used qualitatively to investigate the effect of the wetting agent.

It will be noted that, in general, the greatest reductions in corrosion as brought about by the wetting agent were in the cases of the lowest degrees of wetting (the larger contact angles) of the steel by the inhibited hydrochloric-acid solutions. With the thiourea and substituted-thiourea compounds where the contact angles are fairly high, 32 to 54°, the de-

crease in corrosion rate is about 50 per cent. When the contact angles are fairly small, as in the case of the picolines, the improvement is likewise smaller.

The reason for this reduction in corrosion is not known definitely, but is believed due, in part, to the fact that in the presence of the wetting agent, the hydrogen bubbles have only a very slight contact with the metal surface. As a result of this they do not adhere to the metal

TABLE 1. EFFECT OF WETTING AGENTS ON STEEL CORROSION IN ACIDIZING

(Corrosion Rates Obtained at 79.4°C, on ASTM A10–39 Mild Steel in 2.5% Hydrochloric Acid; Surface Tensions Determined at Acid-Air Interface 25°C, and Contact Angles Formed by Hydrogen Bubbles at 25°C, on Steel Surface in Acid Solutions; Wetting Agent was a Fatty-Alcohol Sulfate of C_8-C_{10} Chain)

Acid Inhibitor	Results Without Wetting Agent			Results With Wetting Agent	
	Corrosion Rate (lb/sq ft/day)	Surface Tension (dynes/cm)	Contact Angle°	Corrosion Rate (lb/sq ft/day)	Reduction in Corrosion (%)
0.1% Pyridine	0.528	67.3	15	0.448	15
0.1% 2-Picoline	0.432	65.6	19	0.326	24
0.1% 3-Picoline	0.504	66.6	16	0.398	21
0.1% 4-Picoline	0.519	68.4	12	0.409	22
0.1% 2,4-Lutidine	0.355	68.6	35	0.217	39
0.1% 2,6-Lutidine	0.282	68.8	32	0.163	42
0.1% 2,4,6-Collidine	0.275	67.3	36	0.162	41
0.1% Quinoline	0.268	64.1	22	0.142	47
0.1% Quinaldine	0.184	65.6	24	0.089	51
0.1% Lepidine	0.261	64.1	24	0.137	48
0.1% Acridine	0.231	67.7	39	0.159	31
0.05% Thiourea	0.325	64.2	32	0.165	49
0.05% Diethylthiourea	0.078	58.6	48	0.038	51
0.05% Dibutylthiourea	0.027	57.0	54	0.013	52
0.05% Diphenylthiourea	0.049	62.1	43	0.024	51

surface, but escape from it in the form of very fine bubbles. Whereas in the absence of a wetting agent, and especially in the cases of the large contact angles, the hydrogen bubbles adhere to the metal surface. This brings about an increase in corrosion which is of a pitting type.

Surface-Active Agents. During acidizing of an oil- or gas well the acid must penetrate into the small capillary pores of the formation. In order to contact the formation the acid must displace the oil from the pore, which at times is rather difficult even though many of the formations, especially those composed of sand, have a connate water [140] film adjacent to the formation, with the oil present in the center of the pore. Surface-active agents [27, 34, 40, 45, 130, 132] are used in the acidizing solutions to help

the acid penetrate and contact the formation. It has been found that the following benefits are derived from the use of such materials:

(1) The acid penetrates the formation easier and at lower pressure.
(2) The acid tends to enter the small pores as well as the large ones without excessive channeling.
(3) The neutralized or "spent" acid returns easier and more completely from the formation after treatment.
(4) There is less tendency for emulsions to be produced between the crude oil and the acid or the spent acid.

Since the formations consist of capillary pores of various sizes, and these pores are filled with oil, water, and gas, the surface forces present · are of extreme importance. While sand and limestone are normally hydrophilic,[46, 85] there are polar compounds [116, 136] in crude oil which can alter the formation through adsorption [8, 59] to such an extent that it may even become hydrophobic [4, 9, 111] in nature. Both acidic and basic polar materials such as naphthenic acid and the homologs of quinoline and isoquinoline[6] have been found in petroleum. These materials are adsorbed upon the acidic silica surfaces and upon the basic limestone surfaces of the formation. Since petroleum varies considerably from oil field to oil field as to types and amount of polar materials, there are formations which have different degrees of wetting by water varying from hydrophilic to hydrophobic in nature.

The surface-active materials used in the acid impart to the solutions such properties as low interfacial tensions with oil and small interfacial contact angles between the acid, the oil, and the formation. Surface-active agents of the anionic and nonionic type are being used for this purpose. Such materials enable the acid to penetrate a formation easier and at lower pressure than an acid solution which does not contain a surface-active agent. Since the acid enters the formation at a lower pressure, there is less tendency for the acid to channel and enter only one section of the formation. The surface-active agents also enable the acid to penetrate the small pores as well as the large ones.

It is important during the acidizing process that the acid after it has become neutralized or spent be returned to the well in order to prevent the blocking of the flow of oil from the formation. There is evidence which seems to indicate that the Jamin effect [57, 58, 67-69] retards the return of the spent acid, and in fact this effect may also increase the pressure required to force acid into a formation. The Jamin effect is brought about by the presence of alternate globules of oil and spent acid within a capillary pore. This system of alternate globules of oil and spent acid is not an emulsion. The globules are fairly large and they touch the sides of the pores. In low-pressure wells the Jamin effect may be caused by the

liberation of carbon dioxide gas bubbles within the spent acid column in the capillary pore. Once the Jamin effect has been established, considerable pressure [132] is required to overcome it.

When an interface such as that between water and oil is present in a capillary, the pressure existing across the interface is given by the formula:

$$P = \frac{2\gamma \cos \theta}{R}$$

where:
P = pressure across interface
γ = interfacial tension
θ = contact angle
R = radius of capillary

If a globule of oil is in the capillary which otherwise is filled with water, there are formed two interfaces. The pressure across the two interfaces is as follows:

$$P_{AR} = \frac{2\gamma_A \cos \theta_A}{R_A} - \frac{2\gamma_R \cos \theta_R}{R_R}$$

where:
P_{AR} = pressure across the two interfaces
γ_A = interfacial tension at the water-advancing interface
θ_A = water-advancing contact angle
R_A = radius of capillary at the water-advancing interface
γ_R = interfacial tension at the water-receding interface
θ_R = water-receding contact angle
R_R = radius of capillary at the water-receding interface

A variation between γ_A and γ_R, θ_A and θ_R, or R_A and R_R results in the pressure P_{AR} between the two interfaces being a definite value, not equal to zero. This gives rise to a resistance to flow of the globule, which is the Jamin effect. The greater the number of globules of oil or bubbles of air present, the greater is the pressure required to bring about movement of the fluid column.

Surface-active agents help considerably in reducing or preventing the Jamin effect by lowering the water-advancing contact angle to approach that of the water-receding angle or at times to reduce both contact angles to zero. The lowering of the interfacial tensions by surface-active agents also reduces the Jamin effect, especially in the case of a gas bubble with water on one side and oil on the other. The water-gas surface tension can be lowered considerably by surface-active agents, and this surface tension will approach that of the oil-gas surface tension.

The surface-active agents usually used for overcoming the Jamin effect during the acidizing process are of the anionic and nonionic types. The agents must be stable not only in the acid solution, but also in the spent acid solution which is about equivalent to a 20 per cent calcium chloride solution.

There are many crude oils encountered during acidizing, which will readily emulsify hydrochloric acid and spent hydrochloric-acid solutions. Even in the case of the spent acid with its high concentration of calcium chloride, very stable emulsions are formed. As the acid penetrates into the formation and travels through the interconnected pores containing capillary columns of oil, there is an excellent opportunity for emulsion formation.

The formation of an emulsion during acidizing brings about an increase in the injection pressure required to force the acid into the formation, which is very undesirable. In the case of the spent-acid emulsions, it is very difficult, if not impossible, to return the acid to the well, and even if the spent acid is returned, the emulsion that is obtained is very difficult to break.

Up to the present time all of the emulsions encountered in acidizing are of the type of water in oil. This has been rather fortunate in that the majority of the emulsifying agents available for use in acid solutions are of the type which tend to form oil in water emulsions. Such agents can be used to break, or prevent from forming, emulsions of the water-in-oil type.

The agent used for the prevention of emulsions must be stable in hydrochloric-acid solutions as well as in spent-acid solutions. No one agent has been found that can be used satisfactorily in acidizing all of the formations in the various oil fields. Thus a number of demulsifying agents are being used. Cationic and anionic, as well as nonionic, demulsifying agents are being employed for this purpose.

Care has to be taken in the choice of an agent to be used for the prevention of emulsions, not only from the standpoint of a satisfactory stability and emulsion preventive, but also in light of the other agents that may be present in the acid solution. Since anionic agents are often used in conjunction with the inhibitor and agents of this type are also employed to help the acid penetrate into the formation, it is not always possible to use cationic demulsifying agents. It is well-known that anionic and cationic agents cannot be used together. As more uses for surface-active agents are found in the acidizing process, it becomes increasingly difficult to find sufficient agents that are compatible with one another. For this reason and because they are not adsorbed as much onto the formation, more and more nonionic surface-active agents are being used.

Chemical Plugging

Next to acidizing, the second most important application of colloidal materials for down-hole treatment of oil- and gas wells is in the substances which are used as agents for sealing or plugging the formation.

Among the various applications for colloidal materials for plugging agents may be mentioned the following:

(1) In oil wells it is often necessary to prevent water or gas from entering the well, and in gas wells water has to be shut off.
(2) During acidizing of a well it is often desired to prevent the acid from entering a porous section.

Permanent Plugging. The problem of shutting off water and controlling the gas-oil ratio of wells is extremely important from the economics of production and from the standpoint of conservation of national resources. Many different chemicals have been used for controlling water
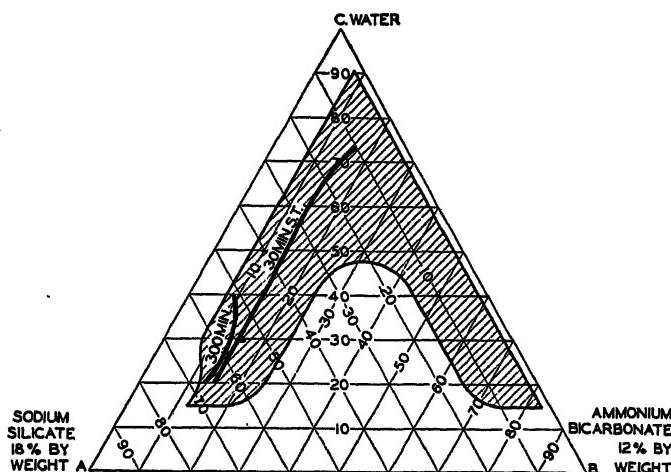
Figure 1. Concentrations of sodium silicate, ammonium bicarbonate and water required to give gels.

and gas entering wells. The colloidal precipitates[74, 75] produced by the brine present in the well and a sodium silicate solution or a sodium carbonate solution have been tried, as well as soaps of various kinds,[36, 50-52, 87] rubber latex,[65] sodium silicate and sodium aluminate,[135] antimony trichloride,[76] silicon tetrachloride,[77] sulfur and aniline,[2] paraffin wax and alcohol,[3] cellulose materials,[88] and various metallic oxides.[86]

Other substances which have been used widely were the gels[83, 117, 118] prepared from sodium silicate and ammonium bicarbonate or from sodium silicate and sulfuric acid. By careful control of the concentrations of these materials[44] as is shown in Figure 1, it is possible to prepare a liquid which can be forced into a water section of a formation where, after a predetermined length of time the liquid will set to a solid silicic-acid gel. The two curves within the shaded area give the concentrations of the materials required for the preparation of the gels having 30- and

300-minutes setting times. Such gels will prevent water from entering a well and considerable success was had using these gels for water control.

The silicic-acid gel has been replaced recently by liquid phenolic-aldehyde resins,[56, 66, 84] which while in the liquid state, can be forced into the formation where they set to hard solids.[7, 21, 98, 113, 138] The liquid resins used to prevent the encroachment of undesirable water or gas are very similar to the casting resins used in the plastic industry.

The liquid resins prepared from the various phenolic-aldehydes materials have properties which make them suitable for water and gas control work. Among these desirable properties are: liquids of good stability can be prepared; the shrinkage of the liquid as it condenses into a solid is not appreciable; the specific gravity of the resins are greater than the majority of the oil-well brines; the resins are insoluble in crude oil and in brines. Liquid resins of sufficiently low viscosities can be prepared, which can be forced into the formations. It is possible, by the use of catalysts and hardeners, to control the time required for the resins to solidify from a few minutes to twenty hours for the range of temperature of 50 to 300°F, which range covers the majority of the oil wells. The usable temperature range of a liquid resin is about 60 to 100 degrees, thus it is necessary to have a series of resins to cover the complete range found in oil wells. The liquid resins which are placed in a well must be either neutral or alkaline because many formations are limestone and dolomite.

The various resins which have to be used are prepared from different mixtures of aldehyde and phenolic materials. The aldehydes usually are either formaldehyde or paraformaldehyde. In the case of the phenolic materials, various ones are incorporated into the resin in order to give it the desired properties.

The reactivity towards the aldehyde of the phenolic material determines the temperature range over which the resin can be used. The more reactive the phenolic, the lower the temperature at which the resin can be employed. The rate of condensation[112] of different phenols with formaldehyde decreases in the following order: resorcinol; 3,5-xylenol; m-cresol; 2,3,5-trimethylphenol; phenol; 3,4-xylenol; 2,5-xylenol; p-cresol; saligenin, o-cresol; and 2,6-xylenol. Figure 2 illustrates how different phenolics can be used to cover various temperature ranges.

The molecular weight upon which the viscosity of the resin depends, can be controlled by the cooking time and temperature used in the manufacturing process. Besides this the cross-linkage of the molecules within the resin also plays an important part in the viscosity. The long-chain molecules are formed from the materials that have two-reactive positions available such as p-cresol, o-cresol, 3,4-xylenol, and 2,5-xylenol. The

materials resorcinol, 3,5-xylenol, and *m*-cresol have three active positions which bring about cross-linkage within the resin.

The solubility of the liquid resins depends to a considerable degree on the phenols used. Resorcinol resins are very soluble in brines, and to some extent the phenol resins are also, especially when large amounts of alkaline catalysts are used. Resins from these materials are not soluble in crude oil. The cresols and the xylenols are brine-insoluble and also crude oil-insoluble whereas the higher alkylated phenols give oil-soluble resins.
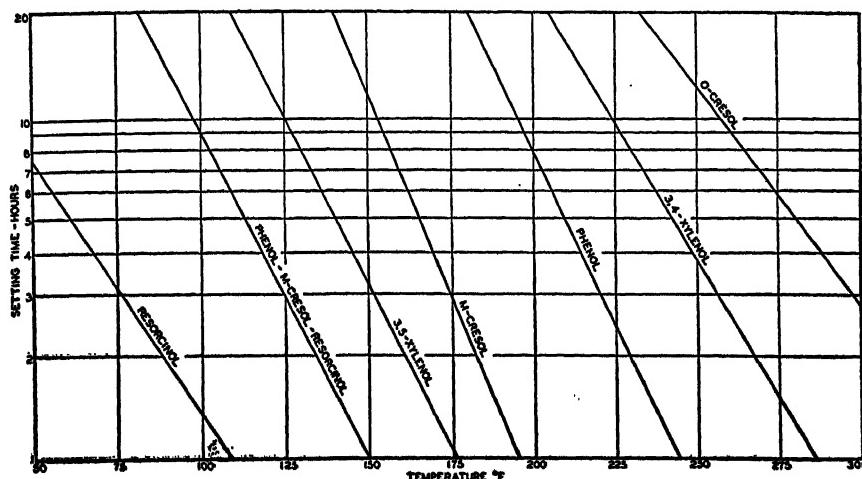
Figure 2. Temperature range of various phenolicformaldehyde resins.

The majority of the phenols when condensed with formaldehyde give resins having a specific gravity of 1.2, which is higher than most brines. In case of the phenols such as the xylenols and higher alkylated materials that do not yield resins of this specific gravity, it usually is possible to incorporate one of the chlorophenols to give the desired weight.

In order to obtain liquid resins that possess all of these required properties, it is necessary to blend several of the phenols, and at times as many as four different phenols have to be used.

Temporary Plugging. Many wells are encountered in acidizing that have a porous zone. Upon acidizing, this section will take all of the acid, with the usual result that very little, if any, increase in oil production is obtained. Thus, at times it is desirable to plug off this zone to prevent the acid from entering it, but at the same time the section can not be permanently plugged if additional production is desired from it. The procedure in acidizing wells of this type is to precede the acid into the well with a plugging agent which blocks off this section to the penetra-

tion of acid, but following the acidizing process the material will, in a relatively short period of time, decompose and reopen the porous section for production. Agents which are used for this purpose are gels of starches and proteins[53] to which have been added specific kinds of bacteria. These gels can be forced into the porous zone and seal it against the entrance of acid.[89, 90] The bacteria, in a predetermined period of time, will decompose and liquefy the gel, thus removing the plugging agent from the well.

Formation Consolidation

Within the last few years liquid resins have been used to consolidate incompetent formations.[18, 129, 140] There are many oil fields in which the sand grains are either not cemented together or are very weakly cemented together. When wells are drilled into such formations and produced, the sand moves into the well, bringing about a reduction and at times a complete halt to the flow of oil or gas. In rare cases a well may produce only a small amount of sand and continue to flow or pump, the sand being carried to the surface in the fluid stream. However, in such instances, the sand erodes the tubing and pumping equipment, and in a short time this equipment has to be replaced.

The problem of formation consolidation is somewhat different from that discussed earlier concerning the use of liquid resins for sealing off undesirable entrance of water and gas into wells. In this case the objective is to cement the sand grains together, at the same time leaving the permeability for the production of oil.

The resins for consolidating formations are somewhat similar to the ones employed for sealing formations, that is, of the phenolic-aldehyde type. As in the case of sealing resins, the choice of phenols depends upon the temperature at which the resins are to be used, the desired setting time of the resins, the viscosity, the specific gravity, and the solubility in brines and crude oil. Here again, various phenols are blended to give the desired properties to the liquid resin.

There are two other important properties which must be incorporated into these resins in addition to those mentioned above. The first property is one of wetting. The resin must preferentially wet the sand grains which are covered with either water or oil. The other property is that the resins when set to a hard solid must coat the sand grains and seal them together, but at the same time must leave permeability within the formation.

In order to make the liquid resins preferentially wet the formations, wetting agents are usually used.[139] Wetting agents employed for this

purpose are long-chain aliphatic amines, sulfonated higher alcohols, and sulfonated naphthenic acids.

In order for the resins to cement the sand grains together and at the same time leave permeability, it is necessary that the resin, while in the liquid state, occupy considerably greater volume than the solid which is formed upon solidification. This is accomplished by having liquid resins of low solid content. By varying the solid content it is possible to control the amount of resin left within the formation,[18] which governs the permeability.

When the liquid resin undergoes condensation to form a solid, it is important that the resin becomes bonded to the sand grains rather than
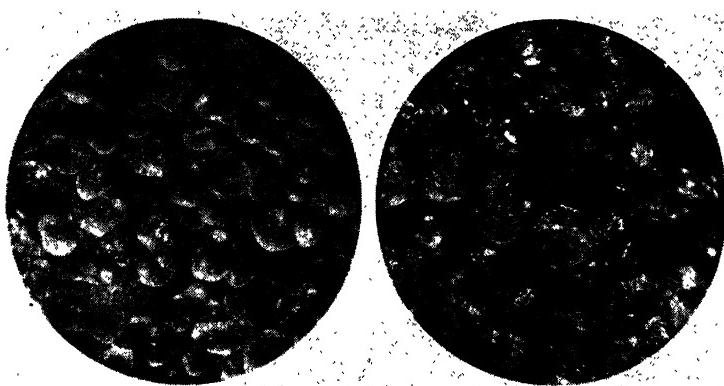
Figure 3. Microphotograph at 10x of unconsolidated sand (*left*) and resin-consolidated sand (*right*).

merely solidifying within the interstices of the sand. This means that the adhesive forces between the solid resin and the sand must be sufficiently strong to give a resin a good bond to the sand, and these forces must be somewhat greater than the cohesive forces between the molecules of the set resin. It is important that this be true because otherwise the resin will exist as a hard solid in the interstices between the sand grains. The cohesive forces existing between the molecules of solid resin are of sufficient magnitude to hold together the sand particles, because once the resin has covered and bonded to the sand grains, the grains are held together only by the cohesive forces with the set resin. The micrograph in Figure 3 illustrates an oil-well sand before and after consolidation with resin.

When these consolidating resins are forced into an incompetent sand formation, the resins first saturate the section in the intermediate vicinity of the well bore. Following this the resin sets to a gel and then begins to shrink onto the sand surface. As a result of this type of shrinkage,

the sand grains are cemented together and there is permeability left in the consolidated section.

Water Flooding

In order to obtain as much oil as possible, more and more fields are being water flooded following the primary recovery methods. This secondary recovery by water flooding has brought about considerable quantities of increased oil production. The colloidal aspects of down-hole treatment of secondary recovery water may be divided into the following topics:

(1) The use of brines in place of fresh water.
(2) The employment of low-pH water.
(3) The surface plugging of the input wells.
(4) The use of surface-active agents.

Use of Brines. It has been known for some time that brines are superior [60, 61, 64] to fresh water for water flooding. To understand why this is true, various studies have been made on the clays found in oil wells in order to determine the factors that influence the rate at which a formation will take water.

The clays usually found in the oil-bearing formations can be classified into three groups, namely, montmorillonite, kaolinite and illite. Three related properties [62] of these clays seem to influence the rate at which an oil-bearing formation will take up water. These properties are (1) base exchange, (2) adsorption and retention of water, and (3) deflocculation and flocculation. Clays of the montmorillonite group possess these properties to a very marked degree, the illite group clays to a somewhat lesser degree, and the kaolinites to a much lesser degree.

The clays as they exist in an oil-bearing formation are swollen to some extent and have adsorbed and retained a certain amount of water. Before any water has been forced through a formation, the clays therein are in contact with the surrounding aqueous conant water media. Base exchange has taken place and the ions present in aqueous media are at equilibrium [63] with those adsorbed on the clay. Whenever an aqueous solution is forced through a formation containing clays, there will be a base exchange [5] taking place which will have an effect on the flocculation or deflocculation of the clays. A partial flocculation of the clays will bring about an increase of the permeability of the formation; thus, the formation will take more water. Likewise a deflocculation of the clays will decrease the permeability and also the rate at which water can be forced into the formation.

If a brine having the same constituents in the same concentrations

as the substances in the aqueous media surrounding the clay particle, is forced through the formation, there will be no change in the swelling of the clays. Such a flood water will have no effect upon the clay particles. Whenever the flood water contains a higher concentration of magnesium, calcium, strontium, barium, and hydrogen ions than is found in the media surrounding the clay, then, by means of base exchange some of these ions will be taken up by the clay. This will bring about a decrease in the swelling of the clay, and thus an increase in the permeability of the formation. Likewise, if the flooding water contains a lower concentration of the divalent metals and a higher concentration of sodium and potassium than does the water surrounding the clay particle, the clay will swell a small amount, resulting in a decrease in the permeability of the formation. Thus, brines are better than fresh water for flooding purposes because there is less likelihood for a brine to cause the clays to swell.[99]

Use of Low-pH Water. It is often necessary to use fresh water as the flooding fluid and in these cases it has been found desirable to use low-pH water.[16, 99] Since the isoelectric points of the clays are at about a pH of 2 to 3, this means that by using fresh water of a pH in this neighborhood the clays will be at the state of minimum swelling. Using low-pH water, it has been possible to force considerably larger quantities of water into the injection well. This usually results in an increase in oil production. It is necessary to use corrosion inhibitors along with water of this pH in order to protect the steel from acid attack. The inhibitors suggested for this purpose are the amine-type wetting agents [14, 15] such as the acetate of n-primary octadecylamine, 1-hydroxyethyl 2-heptadecenyl glyoxalidine, and the hydrochloric of an amine from abietic acid

Surface Plugging. After the water flood projects have been in operation for a period of time, it is often necessary that the oil-depleted zones be blocked off to prevent the input water from by-passing through these sections to the output well. Various colloidal materials have been placed in such wells to plug selectively the surface of the formation of these flooded-out sections. Among the materials which have been suggested for this purpose are wax distillates,[128, 141] wax emulsions,[141] silica gels,[80] urea-formaldehyde liquid resins,[80] colloidal clays,[141] metallic oxides,[134] and colloidal dispersions of solids [109, 110, 141] in air or gas, such as dusts and smokes of rosin, stearic acid, paraffin, and ammonium chloride.

The material which seems to have given the best field results is a resin known as "Dresinol." [71, 94] The dispersion, containing 45 per cent solids having an acid number of 80, is stabilized by means of a protective colloid. The resin is insoluble in aliphatic hydrocarbons and partially soluble in aromatic hydrocarbons. The solid material has an average particle size of less than one micron and can be used to surface-plug

the section of the formation which is taking the greater portion of the water.

   Surface-Active Agents. There are many oil fields which do not respond too satisfactorily to water flooding. In fact, sometimes even after the most effective floods, 25 to 40 per cent of the original oil content is left within the formation. From studies of wettability of silica surfaces by different crude oils and oil-well brines it has been found that it is not uncommon to have an oil which is not displaced by water from silica.[73] Interfacial contact angles upon silica surface have been reported for eight different crude oils and brine solutions.[10] The water-advancing interfacial contact angles varied from 77 to 165 degrees, with only one oil giving a value less than 90 degrees. This indicates that, of these oils on dry silica surfaces, only one will be spontaneously displaced by brine. In order to move the other seven oils, some other driving force has to be used, such as a hydrostatic head on the water.

   The interfacial receding contact angles for the eight oils and brines were all less than 90 degrees and varied from 23 to 50 degrees. This hysteresis effect of the contact angle seems to indicate, as has been mentioned earlier, that polar materials are being adsorbed onto the silica surfaces from the oil, making the solids hydrophobic to the advancing water phase.

   In order to improve the wetting of the formations by brines, surface-active agents have been tried. It has been found that the addition of certain surface-active chemicals to the flood water will increase slightly the oil production and at the same time reduce the residual-oil content of the formation. Water-soluble[133] as well as oil-soluble surface-active agents[17, 133] have been tried for this purpose. These materials will reduce the interfacial tension between oil and water, and also bring about a better wetting of the capillary pores of the formation by water. However, the use of surface-active agents thus far has not been practical because the small increase in oil does not justify the cost of the material. The disadvantage of these materials for this purpose is that they become adsorbed onto the surface of the formation and the flood water becomes depleted of the agent as it passes through the formation. This seems to be true even in the case of the nonionic agents.

Miscellaneous

   There are other applications of colloidal chemicals to down-hole treatment of oil wells, which should be mentioned here although they do not involve colloidal chemistry to as great an extent as the materials discussed heretofore.

In the disposal of oil-well brines back into the earth, sodium septa-phosphate,[24] sodium hexametaphosphate [25] and a quebracho-tannin are often added to prevent the precipitation of calcium carbonate, iron oxides,[115, 127] and barium sulfate.[21, 22]

Many of the fluids produced from wells are corrosive to the steel equipment in the wells. Brines which contain hydrogen sulfide are extremely corrosive, and formaldehyde [23, 35, 95-97] is being used as an inhibitor for this purpose. In gas-condensate wells, where corrosion may be caused by organic and carbonic acids, inhibitors which are semipolar organic compounds [54, 72] are being employed.

In wells producing small amounts of oil, the sand capillaries adjacent to the well bore often become water-blocked and will not allow the passage of oil. This usually means that either (1) the Jamin effect is set up consisting of globules of oil and water or bubbles of gas and water within the capillaries, or (2) the formation becomes blocked by the capillaries saturated with water and there is insufficient pressure to force the oil through the water-filled pores. The production of such wells will be decreased or at times stopped completely. Surface-active agents have been used to remove this water block.[3, 108] The agents are forced into the block, where they will become mixed with the troublesome water. This will bring about a lowering of the interfacial tension between oil and water, allowing the oil to enter the water-blocked capillaries.

Recently a method of fracturing a formation, known as "hydrafrac," [32, 33] has been used to treat oil wells. The fracturing is accomplished by forcing into the formation under high pressure and at considerable velocity an oil gel of Napalm soap. Since the gel is viscous and is injected into the well at a greater rate than it can penetrate into the formation, a fracture is formed. The gel contains suspended sand particles which are forced into the fracture to guard against its closing after the pressure has been released. In order to prevent the gel from permanently plugging the fracture, a small amount of water or a petroleum sulfonate is added to the gel prior to injection into the well so as to convert the gel to a sol after a predetermined length of time. This lowers the viscosity of the Napalm fluid, which is returned to the well.

In the application of cement for sealing the casing and squeezing off of water, it has been reported that the addition of a small amount of bentonite in the order of 1 to 4 per cent imparts to the cement desirable properties [119, 120, 123, 124] such as: elimination of some of the free water caused by settling of cement particles in high-water-ratio neat cement; lowering of the permeability of the set cement; increase in the viscosity of the slurry; increase in the sealing properties of the slurry, thereby preventing some of the formation loss; and production of a set cement which is tougher than neat cement.

Considerable interest along the research line has been given to the idea of making the water-wet producing sand of a well preferentially oil-wet. It is believed by many investigators that a preferentially oil-wet formation will produce a greater quantity of oil than one preferentially water-wet. Many materials have been suggested for this purpose; however today very few actual treatments of oil wells have been completd. Among the various chemicals suggested are: hydrophobic colloids,[101] asphalt,[47] fatty acids,[43] polyvalent metal salt of a fatty acid,[49] crude oil-insoluble bitumen,[48] quinoline, and other organic heterocyclic polycyclic amines.[1]

REFERENCES

1. Albaugh, F. W., U.S. Patent 2,419,755 (1947).
2. Ambrose, H. A., U.S. Patent 2,032,825 (1936).
3. ——, and Teplitz, A. J., U.S. Patent 2,032,826 (1936).
4. Andresen, K. H., "Drilling and Production Practice," p. 447, New York, Am. Petroleum Inst., 1938.
5. ——, and Heck, E. T., "Drilling and Production Practice," p. 236, New York, Am. Petroleum Inst., 1946.
6. Bailey, J. R., "Science of Petroleum," Vol. II, p. 1047, New York, Oxford Univ. Press, 1938.
7. ——, and Dimit, C. E., Oil Gas J., 42 (25), 52 (1943).
8. Bartell, F. E., and Miller, F. L., Ind. Eng. Chem., 20, 738 (1928).
9. Benner, F. C., and Bartell, F. E., "Drilling and Production Practice," p. 341, New York, Am. Petroleum Inst., 1941.
10. ——, Dodd, C. G., and Bartell, F. E., "Drilling and Production Practice," p. 169, New York, Am. Petroleum Inst., 1942.
11. Bockris, J. M., and Conway, B. E., J. Phys. Coll. Chem., 53, 527 (1949).
12. Boner, C. J., in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. VI, p. 553, New York, Reinhold Publishing Corp., 1946.
13. Bonnet, C. F., Oil Weekly, 101 (12), 21 (1941).
14. Breston, J. N., Producers Monthly, 10 (7), 19 (1946).
15. ——, and Barton, K., Producers Monthly, 12 (1), 13 (1947).
16. ——, and Johnson, W. E., Producers Monthly, 9 (12), 19 (1945).
17. ——, and ——, Producers Monthly, 10 (7), 28 (1946).
18. Cardwell, P. H., Petroleum Technol., 10 (2), T.P.2148 (1947).
19. ——, Oil Gas J., 46, 139 (1948).
20. ——, and Eilers, L. H., Ind. Eng. Chem., 40, 1951 (1948).
21. Carroll, D. L., Oil Weekly, 112 (4), 14 (1943).
22. Case, L. C., Oil Gas J., 45 (22), 68 (1946).
23. ——, Petroleum Engr., 19 (3), 150 (1947).
24. ——, U.S. Patent 2,429,593 (1947).
25. ——, U.S. Patent 2,429,594 (1947).
26. Chamberlain, L. C., Oil Weekly, 88 (12), 20 (1948).
27. ——, and Boyer, R. F., Ind. Eng. Chem., 31, 400 (1939).
28. Chapman, M. E., Oil Gas J., 37 (1), 107 (1938).
29. Chappel, E. L., Roetheli, B. E., and McCarthy, B. Y., Ind. Eng. Chem., 20, 582 (1928).
30. Ch'iao, S. J., and Mann, C. A., Ind. Eng. Chem., 39, 910 (1947).
31. Clark, J. B., Trans. Electrochem. Soc., 69, 131 (1936).
32. ——, Oil Gas J., 47 (24), 76 (1948).
33. ——, J. Petroleum Technol., 1 (1), 1 (1949).
34. Clason, C. E., Petroleum Engr., 11 (7), 166 (1940).
35. Clay, J. A., Jr., Petroleum Engr., 18 (2), 111 (1946).

36. Dow, H. H., and Grebe, J. J., U.S. Patent 2,294,078 (1942).

37. Dowell Inc., Tulsa, Okla., *The Acidizer,* 1 (4), (1936).

38. ———, Tulsa, Okla., *The Acidizer,* 2 (1), (1938).

39. ———, Tulsa, Okla., *The Acidizer,* 2 (3), (1938).

40. ———, Tulsa, Okla., *The Acidizer,* 3 (2), (1939).

41. ———, Tulsa, Okla., *The Acidizer,* 4 (3), (1940).

42. Dunstan, A. E., in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. III, p. 491, New York, Chemical Catalog Co. (Reinhold Publishing Corp.), 1931.

43. Eickmeyer, A. G., U.S. Patent 2,452,736 (1948).

44. Fitzgerald, P. E., "Drilling and Production Practice," p. 333, New York, Am. Petroleum Inst., 1939.

45. Flood, H. L., *Petroleum Engr.,* 10 (11), 35 (1939).

46. Garrison, A. D., "Drilling and Production Practice," p. 130, New York, Am. Petroleum Inst., 1935.

47. ———, U.S. Patent 2,241,254 (1941).

48. ———, U.S. Patent 2,246,725 (1941).

49. ———, U.S. Patent 2,354,203 (1944).

50. Grebe, J. J., U.S. Patent 2,090,626 (1937).

51. ———, U.S. Patent 2,152,307 (1939).

52. ———, U.S. Patent 2,152,308 (1939).

53. ———, and Stoesser, S. M., U.S. Patent 1,998,756 (1935).

54. Gross, W. F., and Andress, H. W., *Oil Gas J.,* 47 (26), 76 (1948).

55. Gurwitsch, L., in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. III, p. 523, New York, Chemical Catalog Co. (Reinhold Publishing Corp.), 1931.

56. Hefley, D. G., and Cardwell, P. H., *Petroleum Engr.,* 15 (3), 51 (1943).

57. Herold, S. C., "Analytical Principles of the Production of Oil, Gas and Water from Wells," p. 410, Stanford, Stanford Univ. Press, 1938.

58. ———, "Oil Well Drainage," p. 25, Stanford, Stanford Univ. Press, 1941.

59. Hocott, C. R., *Trans. Am. Inst. Mining Met. Engrs. (Petroleum Development and Technology),* 132, 184 (1939).

60. Hughes, R. V., *Producers Monthly,* 9 (8), 10 (1945).

61. ———, *Producers Monthly,* 10 (11), 9 (1946).

62. ———, *Producers Monthly,* 11 (4), 13 (1946).

63. ———, *Producers Monthly,* 11 (5), 12 (1946).

64. ———, and Pfister, R. J., *Petroleum Technol.,* 10 (2), 1 (1947).

65. Irons, C., U.S. Patent 2,121,036 (1938).

66. ———, and Stoesser, S. M., U.S. Patent 2,274,297 (1942).

67. Jamin, J., *Compt. rend.,* 50, 172 (1860).

68. ———, *Compt. rend.,* 50, 311 (1860).

69. ———, *Compt. rend.,* 50, 385 (1860).

70. Jimeno, E., Grifoll, I., and Morral, F. R., *Trans. Electrochem. Soc.,* 69, 105 (1936).

71. Jones, J. P., *Producers Monthly,* 12 (9), 19 (1948).

72. Justice, W. H., and Jones, E. N., *Oil Gas J.,* 45 (27), 86 (1946).

73. Katz, D. L., *Petroleum Technol.,* 4 (6), 17 (1941).

74. Kennedy, H. T., *Trans. Am. Inst. Mining Met. Engrs. (Petroleum Development and Technology),* 118, 177 (1936).

75. ———, *Oil Gas J.,* 40 (26), 92 (1941).

76. ———, U.S. Patent 2,146,480 (1939).

77. ———, and Lawton, H. C., U.S. Patent 2,019,908 (1935).

78. Langmuir, I., *J. Am. Chem. Soc.,* 38, 2267 (1916).

79. Larsen, D. H., in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. VI, p. 509, New York, Reinhold Publishing Corp., 1946.

80. Lawton, H. C., *Producers Monthly,* 10 (9), 32 (1946).

81. Leach, G. L., *Oil Weekly,* 101 (3), 17 (1941).

82. Lenhard, P. J., *Petroleum Engr.,* 14 (10), 82 (1943).

83. Lenhard, P. J., and Reimers, H. A., *Oil Weekly*, 94 (5), 15 (1939).
84. Leverett, M. C., and Wrightsman, G. G., U.S. Patent 2,366,036 (1944).
85. Livingston, H. K., *Petroleum Engr.*, 10 (4), 84 (1939).
86. Loomis, A. G., U.S. Patent 2,042,011 (1936).
87. ———, U.S. Patent 2,143,991 (1939).
88. ———, Teplitz, A. J., and Ambrose, H. A., U.S. Patent 2,034,347 (1936).
89. Love, W. W., *Petroleum Engr.*, 9 (10), 25 (1938).
90. Lyons, O. W., *Oil Weekly*, 91 (1), 32 (1938).
91. Mann, C. A., *Trans. Electrochem. Soc.*, 69, 115 (1936).
92. ———, Lauer, B. E., and Hultin, C. T., *Ind. Eng. Chem.*, 28, 159 (1936).
93. ———, ———, and ———, *Ind. Eng. Chem.*, 28, 1048 (1936).
94. Martin, D., Andresen, K. H., and Ellenberger, F. W., *Producers Monthly*, 11 (6), 18, (1947).
95. Menaul, P. L., *World Oil*, 127 (6), 150 (1947).
96. ———, U.S. Patent 2,426,318 (1947).
97. ———, and Dunn, T. H., *Petroleum Technol.*, 9 (1), T.P. 1970 (1946).
98. Miller, E. B., Jr., *Oil Weekly*, 114 (3), 65 (1944).
99. Miller, L. E., *Producers Monthly*, 11 (1), 35 (1946).
100. Monson, L. T., and Stenzel, R. W., in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. VI, p. 535, New York, Reinhold Publishing Corp., 1946.
101. Moore, T. V., and Cannon, G. E., U.S. Patent 2,345,713 (1944).
102. Morian, S. C., *Petroleum Engr.*, 12 (7), 67 (1941).
103. Morrell, J. C., and Egloff, G., in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. III, p. 503, New York, Chemical Catalog Co. (Reinhold Publishing Corp.), 1931.
104. Munger, C. G., *Trans. Electrochem. Soc.*, 69, 85 (1936).
105. Muskat, I. E., U.S. Patent 2,010,409 (1935).
106. ———, U.S. Patent 2,016,118 (1935).
107. Nabuco, De Araujo, C. E., Jr., in "Colloid Chemistry, Theoretical and Applied, ed. J. Alexander, Vol. VI, p. 531, New York, Reinhold Publishing Corp., 1946.
108. Newton, W. M., and Jones, P. H., U.S. Patent 2,342,424 (1944).
109. Nielson, R. F., *Producers Monthly*, 12 (2), 15 (1947).
110. ———, and Yuster, S. T., *Producers Monthly*, 10 (1), 16 (1945).
111. Nutting, P. G., "Problems of Petroleum Geology," p. 825, Tulsa, Okla., American Assoc. of Petroleum Geologists, 1934.
112. Powers, P. O., "Synthetic Resins and Rubbers," p. 68, New York, John Wiley & Sons, Inc., 1943.
113. Pryor, C. C., *Petroleum Engr.*, 15 (13), 72 (1944).
114. Putman, S. W., and Fry, W. A., *Ind. Eng. Chem.*, 26, 921 (1934).
115. Redman, F. T., and Mackie, R. F., *Producers Monthly*, 4 (6), 12 (1940).
116. Reid, E. E., "Science of Petroleum," Vol. II, p. 1033, New York, Oxford Univ. Press, 1938.
117. Reimers, H. A., U.S. Patent 2,207,759 (1940).
118. ———, U.S. Patent 2,330,145 (1943).
119. Reynolds, J. T., "Drilling and Production Practice," p. 130, New York, Am. Petroleum Inst., 1942.
120. ———, *Producers Monthly*, 6 (6), 29 (1942).
121. Rhodes, F. H., and Kuhn, W. E., *Ind. Eng. Chem.*, 21, 1066 (1929).
122. Robinson, W. W., *Oil Gas J.*, 38 (46), 190 (1940).
123. Sawdon, W. A., *Petroleum Engr.*, 10 (6), 33 (1939).
124. ———, *Petroleum Engr.*, 15 (4), 152 (1944).
125. Schilthuis, R. J., *Trans. Am. Inst. Mining Met. Engrs. (Petroleum Development and Technology)*, 127, 199 (1938).
126. Shelley, P. G., Stone, J. B., and Coburn, L. W., *Petroleum Engr.*, 12 (10), 60 (1941).
127. Simons, H. F., *Oil Gas J.*, 40 (49), 29 (1942).
128. ———, *Oil Gas J.*, 42 (1), 47 (1943).

129. Smith, R. H., and Polk, A. C., *Petroleum Technol.*, **10** (2), T.P. 2147 (1947).
130. Stone, J. B., and Burge, L. L., *Oil Gas J.*, **40** (6), 42 (1941).
131. ———, and Hefley, D. G., *Oil Weekly*, **99** (10), 32 (1940).
132. Stormont, D. H., *Oil Gas J.*, **38** (8), 33 (1939).
133. Terwilliger, P. L., and Yuster, S. T., *Producers Monthly*, **11** (1), 42 (1946).
134. Torrey, P. D., *Oil Weekly*, **94** (4), 30 (1939).
135. Vail, J. G., U.S. Patent 2,131,338 (1938).
136. Von Braun, J., "Science of Petroleum," Vol. **II**, p. 1007, New York, Oxford Univ. Press, 1938.
137. Weber, G., *Oil Gas J.*, **39** (41), 40 (1941).
138. ———, *Oil Gas J.*, **45** (5), 80 (1946).
139. Wrightsman, G. G., and Buckley, S. E., U.S. Patent 2,378,817 (1945).
140. ———, and Spain, H. H., *Oil Gas J.*, **46** (11), 73 (1947).
141. Yuster, S. T., and Calhoun, J. C., Jr., *Producers Monthly*, **8** (7), 31 (1944).

THE FLUID CATALYTIC CRACKING PROCESS

W. J. Mattox

Esso Laboratories, Baton Rouge, La.

THE PRIMARY PURPOSE of the fluid catalytic cracking process, as with other petroleum cracking operations, is to break down high-boiling, high molecular-weight hydrocarbons into molecules boiling within the gasoline range. The term *fluid* is applied to this type of cracking because the finely divided catalyst is suspended in gas or vapor so that it flows through the equipment with flow characteristics resembling a liquid. Catalyst circulation is maintained without mechanical means. This process, as well as other catalytic cracking processes, has come into existence because of its ability to produce high yields of high quality gasoline. Other widely used modifications of catalytic cracking include the Houdry (static-bed) and the Thermofor (moving-bed) processes. The fluid process differs from these other catalytic cracking methods by certain inherent characteristics which will be discussed in the following sections.[17, 34, 35]

While the gasoline yield advantage of catalytic cracking is important, the potential conservation of our oil supplies through more efficient fuels is of far greater importance. Gasoline produced by the fluid catalytic cracking process has a research octane number of about 100 when tetraethyl lead is added. This fuel will meet the requirements of a 10 to 1 compression ratio engine. In cars having comparable acceleration, hill climbing ability, and other performance features, such an engine is capable of giving about 30 per cent more miles per gallon than our present engines of about 6.6 compression ratio. When used in cars properly designed to take advantage of its high quality, the gasoline produced by catalytic cracking makes possible a decided saving in fuel consumption.[24]

Historically, the first commercial application of catalytic cracking was the Houdry process, in which the catalyst is used in the form of granular particles in a fixed bed. Work on the development of fluid catalytic cracking started well before World War II, when a process was particularly needed to meet the increased demands for motor gasolines of higher

octane number. The Standard Oil Development Company pioneered an investigation of the possibilities of using the catalyst as a powder that could be carried through the reactor as a suspension in the oil vapor to be cracked. The used catalyst could then be passed in a continuous operation through a regenerator where the carbonaceous deposit was burned off. Further intensive pilot-plant work in equipment varying from small laboratory units to a semi-plant scale unit of 100 barrels per day feed capacity led, in a comparatively short time, to full commercial application. The first commercial plant was designed by the Standard Oil Development Company and was placed in operation at the Baton Rouge refinery of the Standard Oil Company of Louisiana (now Louisiana Division of Esso Standard Oil Company) in the early part of 1942.[23, 26, 28]

The large-scale development and adoption of this catalytic process is one of the most impressive and dramatic achievements of the refining division of the oil industry. The combined capacity of the fluid catalytic cracking plants in operation or projected for construction is about 1,400,000 barrels per day, representing some 68 units, or more than half the total catalytic cracking capacity. Capacity of the various units ranges from about 2000 to 41,000 barrels per day. This development makes the petroleum industry the nation's leading consumer of catalysts, a unique position when it is considered that 20 years ago catalytic processes were generally held by the petroleum industry to be research curiosities of potential, but not near-future, importance. Replacement requirements for synthetic catalysts alone have been estimated at 38,000 tons for 1948, these catalysts being valued at $10,000,000.[32]

Feed Stocks

The petroleum feed stocks used for gasoline manufacture in commercial catalytic cracking operations may range from naphtha cuts to heavy gas oils, deasphalted residuums, or reduced crudes. The usual feed, however, is a petroleum gas oil, the most volatile fractions of which have a higher boiling point than the end boiling point of the gasoline produced. Thus, for aviation gasoline production, the initial boiling point of the feed stock will generally be above 300°F; for motor gasoline, above 400°F.

These feed stocks consist essentially of hydrocarbons of the general classifications: paraffins, naphthenes, and aromatics, in various proportions, depending on the nature and the boiling point. The naphthenes and aromatics usually contain more or less long paraffinic side chains. Polynuclear naphthenes and aromatics and these types combined in one molecule to form a naphtheno-aromatic hydrocarbon are widely dis-

tributed in high-boiling fractions. The unsaturated hydrocarbons are almost entirely absent in straight-run products, but they are easily formed in the first stages of cracking and play an important part in cracking and other reactions. The unsaturated hydrocarbons formed in cracking belong mostly to open-chain monoolefins. Cycloolefins and diolefins are of secondary importance.

In addition to the hydrocarbons, petroleum stocks may contain oxygen, sulfur, and nitrogen compounds, including naphthenic and other acids, oxygen-containing neutral resins and asphaltenes, and nitrogen organic bases. The combined quantity of all these compounds does not usually exceed 0.5 to 1.0 per cent in the low-boiling fractions, including gas oils, but may be as high as about 10 per cent in some of the residues produced from heavy asphaltic crudes.

Catalysts for Fluid Cracking

Catalytic cracking is a heterogeneous type of catalysis in which the catalyst forms a solid phase immiscible with liquid or gaseous reactants. Since the reaction takes place on the surface of the catalyst, the nature of this surface is of the utmost importance. The surface area of active catalysts is very large, and usually exceeds 200 square meters per gram when the catalyst is fresh. The extensive surface area is associated with the structure of the material which is actually a network of capillaries and pores. The average pore diameter varies from about 30 to 110 Angstrom units. This porous structure and a high specific surface are common to all the catalysts of this type. An adsorbent having a large surface area is not, however, an active cracking catalyst unless the surface is catalytically specific and activates the adsorbed molecule to be cracked. Some silica gels, for example, with surface areas above 300 square meters per gram have no significant cracking activity. For catalysts of the same type and composition, the activity is proportional to the surface area. Thus, other properties related to surface area may also be used for evaluating activity, such as the adsorption of aromatics and the heat of wetting.[22, 31]

Owing to the nature of the fluid operation, the catalysts employed must be characterized by certain unique properties in addition to those which contribute to their ability to promote the cracking reaction. Of these additional requirements, those of particle size and strength are of special importance. Whereas the strength and density of a pelleted or extruded catalyst intended for use in a fixed-bed or moving-bed type plant may be controlled to a considerable extent by the conditions under which the material is formed, this is much more difficult for the fine powders to be used in a fluid unit. Strength and density must be controlled,

therefore, at some other point in the catalyst manufacturing process, and it has been found that with the synthetic catalysts these properties can be altered in a variety of ways, such as varying the concentrations of the solutions used in making the silica gel, the time the gel is allowed to age before drying, the temperatures of the solutions, etc. With natural clay catalysts also, some control of these properties can be effected during the activation and processing treatment.

The use of catalytic materials of this general type in various modifications of cracking has been known for some time and, although most of the above factors were known to influence the performance of the catalyst, the lack of an adequate understanding of the mechanism of catalytic cracking has necessitated a development, for the most part, on a purely experimental basis.

Up to the present time, only two basically different types of catalyst have been employed in commercial fluid units. These are (1) the natural clay-type catalysts, such as Super Filtrol, and (2) synthetic gel-type silica base catalysts, specifically, silica-alumina and silica-magnesia.

Natural Clay Type Catalysts.[9, 42] These catalysts are chemically-treated, naturally-occurring clays, and in general they cost appreciably less than the synthetic gel-type compositions. Super Filtrol, which is one of the best known and most widely used of the natural catalysts, is an acid-treated bentonite clay mineral, montmorillonite. The pure mineral is ideally of the form $4SiO_2 \cdot Al_2O_3 \cdot H_2O$. The naturally-occurring material frequently has some of the alumina replaced by other metal oxides such as CaO, MgO, and Fe_2O_3. The acid-leached and activated material available commercially usually contains most of these substances in small quantities.

X-ray diffraction patterns indicate that Super Filtrol is composed principally of crystalline montmorillonite. This material is characterized by an excellent activity for cracking heavy oils and has a high surface area (approximately 300 square meters per gram). It is quite stable up to temperatures of about 1300 or 1400°F, but it undergoes profound changes in surface area, crystalline structure, and catalytic activity when heated to higher temperatures—in the range of 1400 to 1500°F or higher. At these very high temperatures, which are well above those employed in the cracking process, Super Filtrol's cracking activity vanishes for all practical purposes, its surface area drops to a very low value, and its crystalline structure changes to an inactive aluminum silicate.

Synthetic Silica-Base Catalysts.[2, 6, 8, 15, 41, 42] Silica-alumina composites have been widely used commercially in the fluid cracking process. These materials are characterized by an amorphous structure as indicated by x-ray diffraction patterns and a high surface area. The composition normally ranges from 85 to 90 per cent silica and from 10 to 15 per cent alumina. The most commonly found impurities are generally present

in very small amounts, since these materials may have a marked influence upon the catalytic properties. Alkali metal compounds such as Na_2O are also usually present in trace quantities.

A number of techniques of preparing these catalysts are known, among which are included the three following methods: (1) precipitation in sequence, in which alumina is precipitated from an aluminum salt solution in the presence of a slurry of hydrous silica gel, (2) mixing of the two wet gels in appropriate proportions, and (3) impregnation of partially dried silica gel with an aluminum salt and subsequent thermal decomposition to form alumina. The precipitated or treated gels are washed and dried, and they may be calcined at temperatures of the order of 1000°F. If the catalyst is calcined, most of the water is removed, but a small amount (up to about 2 per cent) remains and is an important part of the active catalyst. It is believed that this *constitutional* water forms active hydrogen atoms of an ionic nature, that are associated with catalytic activity.

The fresh silica-alumina catalysts normally show specific surface areas as high as 400 to 600 square meters per gram and rather high pore volumes of about 0.5 cc per gram. The sizes of the pores are small according to ordinary standards and are of the order of 50 Ångstrom units. However, these openings are much larger than the gas oil molecules which must enter them to undergo reaction, or the smaller molecules, formed as a result of the catalytic reaction, which must leave the catalyst.

The silica-alumina catalyst is quite stable to heat, much more so than the natural clays. The synthetic material loses about half of its activity and surface area when heated in atmospheric air to temperatures of the order of 1600 to 1650°F. It is still amorphous after such a treatment. For all practical purposes, however, the material is inactive catalytically and has a very small specific surface after it has been calcined at 1700 to 1800°F. Heat-treatment at higher temperatures causes complete collapse of the structure, resulting in the formation of such crystalline substances as cristobalite, a form of quartz, and mullite or sillimanite, which are aluminum silicates.

Synthetic silica-magnesia [1, 5, 33, 42] appears to be the most promising of other silica-base materials. Higher gasoline yields are achieved as compared to the silica-alumina catalyst, but the gasoline octane number is lower. This catalyst maintains its activity better than silica-alumina.

Silica-magnesia catalysts are similar to silica-alumina composites in that they are characterized by amorphous x-ray diffraction patterns and by a very active type of silica base. The surface area of the magnesia catalyst usually varies from 400 to 600 square meters per gram and the average pore diameter is in the range of 30 Ångstrom units for the fresh material. When heated to about 1400°F, it is converted into a crystalline,

inactive, and relatively nonporous magnesium silicate. Thus, with respect to thermal stability, this composition behaves much more like a montmorillonite clay than the synthetic silica-alumina product.

Basic Elements of the Fluidized Solids Technique [21, 25, 29, 30]

If finely divided solids, such as sand, are placed in a cylinder and air is introduced at the bottom of the cylinder at a very low rate through a porous plate, there is no expansion of the bed of powdered solids and the air flows through the powder in the same way as air would flow through a tower filled with Raschig rings or other packing material. As the velocity of the air is increased, the pressure drop through the bed of powder rises until the entire bed and the individual particles of which it is composed are supported on the rising current of gas. A slight expansion of the bed occurs, corresponding to a lifting of the particles out of contact with each other; the entire bed assumes a high degree of mobility, such as is characteristic of a liquid, and is considered to be fully fluidized. At this point the pressure drop through the bed is essentially equal to the bed weight. Further increases in air velocity do not result in increased pressure drop, but instead the bed is further expanded and its density is decreased; also, increasingly turbulent motion of the individual particles within the bed is induced, and entrainment of solid by the gas leaving the bed may begin.

Particle size is of importance in obtaining a properly fluidized bed. If the size of the particles of powder is too small, the gas tends to channel through fixed flow paths which form in the bed, and the latter is not effectively fluidized. If the size of the particles is too large, there is a tendency for the gas to rise through the bed in excessively large bubbles. Unsteadiness in operation and poor contact of the gas with the solid may result, and in laboratory scale operations "slugging" of the bed may occur; that is, segments of the solid may be trapped between bubbles of gas which essentially fill the cross section of the vessel or tube and may be forced upwards much as a piston is moved by gas pressure. With particles in the right size range, however, uniform fluidization and excellent gas-solid contact can be obtained.

The usual size range of powdered catalyst employed in a fluid cracking process is from about 10 to 150 microns. A rather surprising observation, which represents, in fact, a key point in the experimental basis for the wide application of the fluid solids technique, is that with particles in this size range fluidized beds of high density can be maintained by a very simple technique at gas velocities which are many-fold higher than the free-fall velocities of the particles contained in the bed, as calculated by Stokes' law or its modifications for turbulent settling conditions; in

effect, at conditions such that the formation of a dense bed would have appeared impossible. In practice, it is found that the rate of entrainment from beds of fine particles fluidized at velocities above the free-fall velocity is rather low, so that the bed may be maintained indefinitely under such conditions by the continuous introduction of a relatively small quantity of solid.

Fundamentally the fluid solids technique may be said to represent a useful application of two basic properties or characteristics of finely-divided solids. These properties may be summarized as follows: (1) Properly sized solids, when mixed with a gas, will form a solids-gas mixture or "fluid," this "fluid" having flow properties similar to those of ordinary liquids. (2) A powdered solid when suspended in a gas stream flowing upwards at moderate velocities will form a continuous dense phase which

Figure 1. Basic elements of the fluid-solids technique.
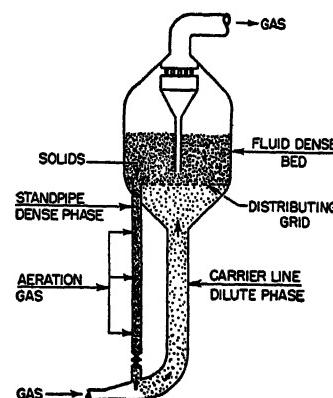
in many aspects resembles a boiling liquid and which assumes a relatively well-defined level.

The first of these properties makes possible the circulation of vast quantities of solids without benefit of mechanical devices, such as pumps, elevators, etc., and it also allows the use of more or less conventionally designed equipment, such as pipes, valves, and exchangers, to handle and control the solids. The second property provides uniform temperatures in exothermic or endothermic reaction beds and supplies means for carrying out such functions as heat transfer, under excellent conditions.

The fluid solids technique as applied to manufacturing processes consists essentially of either one or both of two basic elements, namely, a circulating system and a zone in which a concentrated or "dense phase" of solid is built up or maintained for the purpose of carrying out a desired reaction. An illustration of a simple combination of these two elements is given in Figure 1.

Circulation of Solids. The transfer of fluidized solids in large amounts between two or more vessels is easily carried out by use of a new technique. Such circulation systems consist of (1) a standpipe in which the static head of the flowing solid itself builds up the pressure requisite for its circulation; and (2) a carrier line which contains a flowing stream of gas into which the solids are dispersed, to be carried to a reaction zone. The fluid solids in the standpipe are maintained in a state of relatively high density in order to obtain reasonable build-up of static pressure per unit length of standpipe. However, the solids-gas mixture, being compressible, can attain excessively high densities such that the mixture loses its "fluid" characteristics; gas is accordingly added to the solids-gas mixture at points along the length of the standpipe to maintain its density
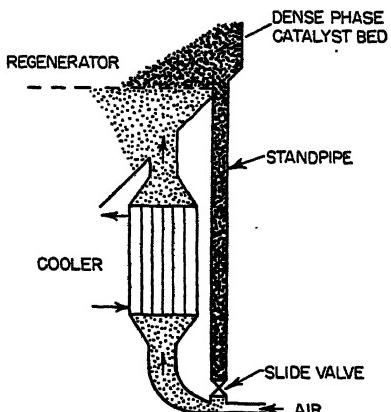
Figure 2. Regenerated catalyst cooler.

in a "fluid" range. Control of catalyst flow in the standpipe is maintained by means of a modified gate-type valve located at the base of the standpipe. Pressure having been built up in this manner, the solids-gas mixture can then be injected into a suitable gas stream for conveyance to any desired point, normally a reaction zone. The density of this solids-gas mixture in the carrier line is maintained low in order to facilitate circulation of the solids. The low-density mixture is obtained simply by operating at relatively high gas velocities in the line, which results in a low ratio of solids to gas in that section.

The Fluid Dense Bed. A zone or bed of high solids concentration can be established and maintained simply by controlling the velocity of a solids-gas mixture at a relatively low level, e.g., 1 to 2 ft/sec, at which conditions the solids tend to settle out and disengage from the gas. The bed formed as a result of this phenomenon is of more or less uniform density, but is in a continuous state of violent agitation—a fact which results in extremely efficient mixing of the solids, and hence in uniform

temperature maintenance. Relatively high densities can be built up in such a bed. For example, with finely divided clay having a settled density of about 45 pounds per cubic foot, bed densities of 15 pounds per cubic foot or higher can readily be obtained. With denser materials such as finely divided iron oxide having a settled density of 150 pounds per cubic foot, bed densities of the order of 125 pounds per cubic foot are readily obtained. In cases where circulation is carried out, solids can be charged to the bed either directly from a standpipe in a dense state, or in a dilute phase by means of the fluidizing gas, as is shown in the drawing (Figure 1). Similarly, removal of the solids from the bed can be accomplished by direct withdrawal through standpipes, or by entrainment overhead with the effluent gases. The former method results in simpler equipment to separate the solids from the gas.

Versatility of Application. The specific arrangement and combination of the basic elements of the fluid solids technique are obviously subject to wide variation, depending on the particular application. For example, functions such as heat removal or exchange can be readily carried out either by immersing coils directly in a dense phase bed, or by recycling solids through a standpipe and heat exchanger located in the dilute phase carrier line (Figure 2). Furthermore, large quantities of heat can readily be transferred from one reaction zone to another because of the very large heat capacity of the circulating solids stream. The fact that the fluidized bed is extremely turbulent insures good mixing and even temperatures throughout the bed, even though the reactions taking place involve high rates of heat liberation or absorption. The fluid technique is obviously versatile in its application, and may be used in a variety of processes which involve both catalytic and noncatalytic reactions and which may or may not require the circulation of solids.

Fluid Cracking Process Flow

A fluid catalytic cracking plant of 25,000 barrels per day capacity is capable of handling more than a million and a quarter tons of feed stock per year. Such units frequently are 250 feet high—as lofty as a twenty-story building—and contain more than 24 miles of piping, of which one section may be as much as 8 feet in diameter. These plants may be considered as comprising four principal sections: feed-preheat section, reaction section, regeneration section, and product-fractionation section. Figure 3, which illustrates one of the more modern types of operation, may be referred to in following the process flow.[18, 23-25, 27, 28]

In the feed-preheat section, fresh charging stock is heated by heat exchange with hot streams from the fractionator and, in some cases, by means of an oil or gas-fired furnace. This preheated feed is directed to

the regenerated catalyst pick-up point where it contacts the hot, pow-
dered catalyst from the regenerator. A rapid transfer of heat takes place
when the hot catalyst joins the preheated oil and the resulting oil vapors
carry the catalytic material as a suspended powder into the reactor.
Owing to the decreased velocity of the vapors passing through the re-
actor, the suspended catalyst particles settle out to form a relatively
dense, turbulent bed. Cracking is effected in this bed at a temperature
of 800 to 1000°F.

Product vapors pass overhead from the catalyst bed through a settling
zone to cyclone separators where all but a trace of catalyst is removed.
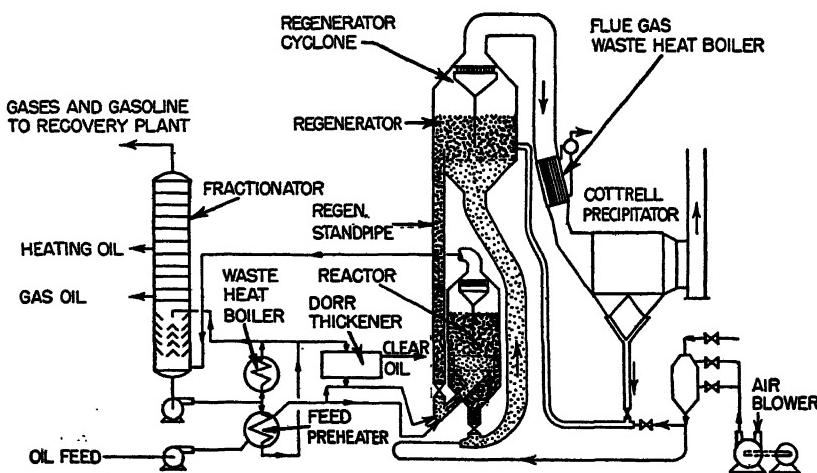
Figure 3. Flow diagram of fluid catalytic cracking plant.

The catalyst separated in the cyclones is returned through a small stand-
pipe to the catalyst bed. The cracked vapors, now substantially catalyst-
free, proceed to a fractionator where they are separated according to
desired boiling ranges, usually into gas, high octane number gasoline,
heating oil, and gas oil fractions. The very small percentage of the
catalyst that escapes from the cyclone separators is recovered as a heavy
oil slurry at the bottom of the fractionator, from which it is pumped
back into the reactor.

During the cracking reaction, the catalyst accumulates a deleterious
deposit of carbonaceous material which must be removed in order that
the catalyst surface be maintained effective for the cracking reaction.
To accomplish this regeneration, carbonized catalyst is continuously
withdrawn from the bottom of the reactor and is directed through a
steam stripping zone where entrained and volatile hydrocarbons are re-
moved. The catalyst is then discharged into an air stream which carries

it in suspended form to the regenerator vessel, wherein a substantial proportion of the carbonaceous deposit is removed by air oxidation. Because of the decreased velocity of the carrying gas, a fluidized bed of catalyst is established within the regenerator, as in the reactor. The regenerator operates at a catalyst bed temperature of about 1100°F— a dull red heat. Most present-day catalytic cracking plants operate in "heat balance"—that is, all the heat released in the regenerator in burning carbon (except heat losses from the equipment) is used to vaporize and heat the oil feed, and to provide the heat of cracking. This heat is transferred from the regenerator as sensible heat in the circulated catalyst stream, and therefore the temperature held in the regenerator can be controlled by the rate at which catalyst is circulated in the system. This catalyst circulation rate must, however, be consistent with other conditions in the reactor to give the desired feed conversion and carbon production.

During the regeneration, sufficient coke is burned from the catalyst to restore its activity and the regenerated material is separated from the spent air (flue gas), first in a settling zone above the bed and then in cyclone separators, after which it is returned through a standpipe into the stream of hot oil feed. At this point the catalyst begins a new journey into the reactor and a fresh cycle of operation.

The flue gas leaving the cyclone separators goes to a waste heat boiler and then to a Cottrell electrical precipitator which removes most of the last traces of catalyst. The gas is vented from the stack of the precipitator and the final recovered catalyst is repressured in a standpipe and passed to the regenerator.

Of particular interest is the fact that the complete cycle takes place without the use of a single mechanical moving part. It is the fluid condition of the catalyst that makes this possible. The average particle of catalyst makes its way around the complete circuit approximately every 10 minutes, yet the process runs continuously for months on end (some runs have lasted considerably longer than a year). In a typical commercial unit processing approximately 25,000 barrels of feed a day, about 20 to 40 tons of catalyst are circulated through the unit every minute, yet the catalyst recovery is so complete that only 0.005 per cent of the 60,000 tons circulated daily is lost to the air.

Catalytic Conversions

Optimum Cracking Conditions.[27, 33, 39, 40] These conditions are achieved in the fluid reactor as a result of the ability to adjust the catalyst holdup, temperature, recycle oil rate, and catalyst rate at the specific values required for the particular charge stock being processed. The reactor of a

fluid unit is so completely flexible that the catalyst bed depth can be adjusted at will to give a wide variation in catalyst holdup, the temperature can be varied from that required to vaporize the feed to over 1000°F, and the proportions of fresh feed and recycle can be varied over the full range of 0 to 100 per cent recycle.

Recycling. Recycling of the higher-boiling cracked products along with the fresh oil feed is frequently employed to produce the maximum yield of more valuable products, such as gasoline and heating oil.

Catalyst-to-Oil Ratios. These vary between a minimum of about 5 to 1 and a maximum of about 20 to 1. In plants where the heat of cracking is supplied by the sensible heat of the circulating catalyst, the catalyst-to-oil ratio is fixed by the severity of cracking and the temperature of the regenerated catalyst.

Spent Catalyst Stripping. Both in initial cost and operating cost, the regenerator is the most expensive single item of equipment, and all reasonable means are used to confine its function to the burning of unavoidable carbon. Provided cracking conditions are optimum, the major source of avoidable carbon results from incomplete stripping of hydrocarbons from the catalyst leaving the reactor. This effect has been minimized by present-day efficient stripper design and by use of the maximum permissible amount of stripping steam.

Reactions in Catalytic Cracking

Although the catalytic cracking of high-boiling petroleum feed stocks represents a quite complicated series of simultaneous and superimposed reactions that vary with the nature of the feed and the cracking conditions, some understanding of the basic reactions involved has been gained through a study of the cracking of a number of pure hydrocarbons. The specific action of cracking catalysts on hydrocarbons depends largely upon the nature of the catalyst and for this reason the reactions discussed here will be in relation to the activated clay and synthetic, silica-base type materials described in a preceding section. These so-called adsorbent-type catalysts moderately and selectively weaken carbon-carbon bonds, particularly those bonds which are located toward the center of the molecule. The formation of hydrogen shows that dehydrogenation reactions are also accelerated by these catalysts, but to a lesser degree. This directional action of the catalyst results in higher yields of valuable liquid products. A number of secondary reactions, such as isomerization, and hydrogen and alkyl transfer, are of importance and will be discussed in connection with the reactions of the individual hydrocarbon types.

Paraffins.[7, 10, 11, 20] The effect of the catalyst on the carbon-carbon bond is highly selective. The third carbon-carbon bond and those nearer the center of the molecule are most strongly affected, with the result that higher yields of the desired lower-boiling liquid hydrocarbons are produced with correspondingly low C_1 and C_2 gas formation as compared to the noncatalytic or thermal process. High yields of C_3 and, particularly, C_4 fractions are also characteristic of the action of these catalysts. The gasoline boiling-range products resulting from the cracking of high molecular-weight normal paraffins contain both branched-chain and straight-chain paraffins and olefins. The normal paraffins are far more stable toward catalytic cracking than the corresponding olefins although the rate of catalytic cracking is from $3(C_3H_8)$ to $60(C_{24}H_{50})$ times greater than that of the noncatalytic process and increases with the molecular weight of the paraffins. The reaction rates may also vary somewhat with the type of catalyst employed. A study of the effects of branching on the catalytic cracking of hexanes has shown that the isomers, with the exception of 2,2-dimethylbutane, are less stable than n-hexane. Thus, the tertiary carbon structure accelerates the catalytic reaction and the quaternary structure retards it, as compared with the normal structure. Paraffins containing a lage number of methyl groups usually form more methane on cracking than the normal paraffins.

Secondary reactions, particularly of the olefins formed, readily take place. In addition to isomerization, transfer of hydrogen takes place on a large scale in the presence of the cracking catalysts, which results in a saturation of olefins and the formation of aromatics. The aromaticity of the fractions formed on catalytic cracking of paraffins increases with increasing molecular weight or boiling point of the product. Some of these highly aromatic fractions contain 50 per cent or more of aromatics. The higher the temperature of the cracking the more aromatics are formed.

Olefins.[4, 7, 12, 37] Olefins are not present in any appreciable quantities in commercial feed stocks. However, the unsaturates formed as a result of the cracking of other hydrocarbon types have a very significant role in the catalytic transformations that result in high-quality gasoline products. These catalytically-formed unsaturates are more difficultly desorbed from the active surface of the catalyst and thereby are subjected to various secondary reactions.

The catalytic cracking of pure, aliphatic olefins shows the same general directional effects as for the paraffins; that is, the splitting reactions effected by the catalyst yield mostly C_3 and larger fragments. The rate of cracking is much faster than for the corresponding paraffin and may be several thousand times faster than that of thermal or noncatalytic cracking. Isomerizations occur to a large extent, both by shifting of

carbon atoms in the skeleton, and by shifting of the position of the double bond of either the original olefin or the lower molecular-weight olefins formed on cracking. Reactions involving hydrogen transfer also have been firmly established. In conversions of this type, the olefin content is greatly reduced and the proportion of aromatics increased; formation of free hydrogen is slight. The formation of paraffins is attributed to the hydrogenation of olefins by hydrogen liberated by the various aromatization reactions and by condensations which result in catalyst deposits of low hydrogen content.

Catalytic cracking of cyclic olefins gives results very similar to those obtained with aliphatic olefins. Isomerization and hydrogen transfer result in the formation of saturated naphthenes, aromatics, catalyst deposits, and, under more severe cracking conditions, polymerization and ring splitting.

Naphthenes.[3, 10, 13] The action of the cracking catalyst on naphthenes results largely in rupture of the naphthenic rings, with formation of saturated and unsaturated fragments. Six-membered naphthenes are dehydrogenated to aromatics only to a small extent. With decalin, the rupture of the rings is predominant, but with a minor amount of dehydrogenation to tetralin and naphthalene. As a result, paraffins (C_3 and C_4), olefins, cyclopentanes, and cyclohexanes are formed with only small amounts of benzene derivatives. Work with tetralin has shown a stronger dehydrogenation tendency amounting to as much as 50 per cent of the total conversion. The remaining portion of the converted tetralin yields benzene and its homologs as a result of the rupture of the naphthenic ring. The cracking of paraffinic side chains is characterized by the same general preferential splitting as established for paraffins, with formation of C_3 and larger fragments. The effect of isomerism on the catalytic cracking of naphthenes may be appreciable and methylcyclopentane is somewhat more susceptible to the action of these catalysts than is cyclohexane at the same conditions. This effect is due to the presence of the tertiary structure in the methylcyclopentane which is particularly susceptible to catalytic cracking, as was shown for the isomeric paraffins.

Work done on the cracking of both monocyclic and bicyclic naphthenes has shown that the catalyst accelerates the rate of cracking about 1000 times as compared to noncatalytic cracking. The rate of catalytic cracking of the naphthenes is several-fold that of paraffins of the same number of cabon atoms. Thus, the order of stability of paraffins and naphthenes is reversed in the presence of the catalyst.

Aromatics.[10, 14, 18, 19, 38] The ring structures of the aromatics are much more stable toward the action of cracking catalysts than are the paraffinic, olefinic, and naphthenic hydrocarbons discussed previously. However, the action of these catalysts on substituted aromatics is one of the most impor-

tant reactions of cracking since it results in the formation of low molecular-weight paraffins, olefins, and aromatics with fewer or shorter side chains. Toluene and diphenyl are quite stable under the usual catalytic cracking conditions. Polymethylated aromatics are much more reactive and are intermediate to paraffins and naphthenes in reactivity. Aromatics with propyl and larger side chains are highly unstable but the initial reaction is principally one of total dealkylation that results in an unsubstituted aromatic and an olefin having the same number of carbon atoms as the alkyl substituent. The reaction of ethylbenzene is very slow under comparable conditions (752 to 932°F). The catalytic reactions effected under much more severe cracking conditions result in total rupture of aromatic rings with formation of hydrogen, methane, coke and other products. Of the isomeric propyl and butyl benzenes, the normal alkyl groups are more stable than the branched ones. The relative decomposition rates of the butyl benzenes increase in the order of normal, secondary, and tertiary. Tertiary butylbenzene is particularly susceptible to catalytic cracking although it has a quaternary structure, which in purely paraffinic hydrocarbons is resistant to catalytic cracking.

The methyl groups in polymethylbenzenes exhibit a unique mobility in the presence of cracking catalysts that may result in isomerization or a disproportionation of these groups between aromatic rings. Under conditions within the range of those usually employed in catalytic cracking, xylene is disproportionated to toluene and trimethylbenzenes, trimethylbenzenes to toluene, xylenes, and other polymethylbenzenes. Under the same conditions (950 to 1000°F), ethyl groups show a greater tendency to dealkylate, and diethylbenzene is converted almost entirely to benzene and ethylene with no disproportionation. This dealkylation reaction, however, has been shown to be reversible. With xylene in the presence of an excess of benzene, disproportionation results primarily in the formation of toluene. Similarly, diethylbenzene and diisopropylbenzene yield the corresponding monoalkylbenzenes.

General. The above discussion of the reactions occurring in the cracking of individual hydrocarbons undoubtedly presents an over-simplified picture of the reactions taking place during the catalytic cracking of a petroleum oil. In the latter case all of the molecular types may be considered to be present, and they and their fragments are known to interreact in a variety of ways so that the overall reaction pattern is exceedingly complicated. It is obvious that a large amount of additional investigation of the interactions of the various molecular species upon each other in the presence of cracking catalysts will be required before a clear understanding of the reactions occurring in the exceedingly complex system of petroleum catalytic cracking can be obtained.

Catalyst Regeneration [36, 37, 40]

The combustion of carbonaceous deposits under closely controlled combustion conditions is an important feature of the fluid cracking process. Regeneration not only restores catalyst activity but contributes all or a major portion of the heat requirements of the unit. The turbulence in the fluid bed of the regenerator gives some remarkable results. In the regenerator of a commercial unit as much as 400,000,000 Btu per hour may be liberated, yet the temperature in the regenerator seldom varies more than 5°F from top to bottom, even though the catalyst bed may be 20 feet deep and up to 55 feet in diameter.

The carbonaceous deposit that is an inevitable by-product of catalytic cracking is usually called *coke* and is generally determined as the weight per cent of carbon on feed or on catalyst. The extent of coke formation depends on the type of catalyst, the feed stock, and the operating conditions. The weight per cent of coke on the catalyst entering the regenerator usually varies between less than 1 to about 3 per cent. A balance between several factors usually determines the coke level desired for any specific operation. First, the burning of a given amount of coke from catalyst having a high deposit generally proceeds more readily than the burning of the same amount from catalyst having a low coke content. Second, it is desirable from the standpoint of catalyst performance during the cracking reaction to maintain a relatively low catalyst carbon content.

Catalyst coke is composed principally of carbon and hydrogen, but may include some sulfur. The actual composition of the deposit is complex as regards hydrocarbon types and is not well known, although the formula $(C_3H_4)n$ has been reported. Depending upon the adequacy of the carbonized catalyst steam-stripping facilities, which are intended to remove entrained and volatile hydrocarbons, leaving only coke, the carbon content of sulfur-free deposit varies from about 93 to 87 weight per cent, and the hydrogen content from about 7 to 13 weight per cent. Since about four times as much air is required to burn one pound of hydrogen as is necessary for one pound of carbon, the importance of adequate catalyst stripping and low hydrogen content is obvious.

The principal factors involved in the regeneration of the fluid catalyst are pressure, temperature, catalyst deposit, catalyst residence time and oxygen concentration in the flue gas.

Regeneration Pressure. Commercial unit regenerators operate with pressures at the top of the vessels ranging between about 1 and 15 pounds per square inch gage. Obviously, with specified maximum allowable gas velocities within the regenerators, the higher pressures permit the use of smaller vessels or increased air flows in existing regenerator vessels.

It has been shown that the combustion of catalyst deposits proceeds more favorably at the higher pressure levels. The choice of the pressure used is based on a consideration of these factors, and others, as compared to the increased cost of compressing large amounts of air to the higher pressures.

Regeneration Temperature. Commercial fluid catalytic cracking units operate with regeneration temperatures ranging from slightly below 1000°F to about 1200°F. The choice of operating temperature usually is a balance between several factors. Since it is desirable to maintain the catalyst-oil ratio at a reasonable minimum, the regeneration temperature is preferably held as high as is consistent with good catalyst life and mechanical limitations of the equipment at the elevated temperatures. In general, the higher the temperature, the more readily the combustion occurs. On the other hand at the highest temperature levels, permanent catalyst deactivation occurs at an accelerated rate, and construction problems and costs tend to become more important considerations. The type of catalyst to be employed also must be considered, since some are more heat-resistant as regards activity decline than others.

Type of Catalyst.[33] As already mentioned, the type of catalyst used affects the operation of regeneration equipment because of temperature deactivation problems. The catalyst type also will affect regeneration-air requirements and heat evolution, since catalysts having different chemical compositions result in various ratios of CO_2 to CO in effluent regeneration gases. With various commercial catalysts the CO_2 to CO ratio varies between 1:1 and 2:1 as shown in the following tabulation.

Catalyst	CO_2/CO Ratio in Regeneration Gas
Natural clay	1.7–2.2
Silica-alumina	1.0
Silica-magnesia	1.6–2.0

The presence of certain metallic contaminants, such as iron, which may accumulate on the catalyst during operation, tends to increase the CO_2/CO ratio, and results in increased air requirements for burning of any given amount of catalyst deposit.

Fluid Cracking of Paraffinic Gas Oil [33]

The data shown in Table 1 give typical results on the cracking of a paraffinic gas oil (29.4°API-701°F mid-boiling point) over the three principal types of commercially available catalysts:

TABLE 1. COMPARISON OF CATALYSTS IN FLUID CATALYTIC CRACKING

All Data at 60% Conversion of Wide-Cut Paraffinic Gas Oil Feed Stock

	Catalyst		
	Silica-Magnesia	Silica-Alumina	Natural Clay
Temperature, °F	975	975	975
Catalyst-to-oil ratio	6–10/1	9–12/1	7–10/1
Yields on feed			
10-lb R.V.P. gasoline, vol. %	56.2	45.5	47.8
Total butanes-butenes, vol. %	10.0	16.0	14.0
Butenes, vol. %	6.4	9.0	9.2
Isobutane, vol. %	2.6	5.4	4.0
Heating oil base, vol. %	22.3	21.5	17.6
Heavy gas oil, vol. %	17.7	18.5	22.4
Dry gas, wt. %	6.2	9.0	8.7
Carbon, wt. %	2.9	2.9	3.1
Gasoline quality			
Motor octane number	79.2	81.6	79.8
Research octane number	91.5	95.0	93.6
Res. oct. no.—leaded, 1.5 cc. TEL/gallon	96.3	98.7	97.2
Volatility, ASTM distillation			
% D+L @ 158°F	25.0	32.0	27.5
% D+L @ 212°F	46.0	54.0	52.0
% D+L @ 257°F	61.0	65.5	64.5
Heating oil base quality			
Gravity, °API	29.5	30.0	28.5
Cetane no. (est.)	36	32	34
Initial b.p., °F	455	470	465
50% Distilled @ °F	514	517	510
Final b.p., °F	577	589	572

REFERENCES

1. Anon., *Chem. Eng.*, **56**, 74 (April, 1949).
2. Baker, M. O., Chesnutt, S. D., and Wier, T. P., Jr., *J. Phys. Coll. Chem.*, **52**, 1364 (1948).
3. Bloch, H. S., and Thomas, C. L., *J. Am. Chem. Soc.*, **66**, 1589 (1944).
4. Ciapetta, F. G., Macuga, J. J., and Leum, L. N., Paper presented at Am. Chem. Soc. Meeting, Atlantic City, N. J., April, 1947.
5. Conn, A. L., Meehan, W. F., and Shankland, R. E., Paper presented before Am. Inst. Chem. Engrs. Regional Meeting, Los Angeles, Calif., March 8, 1949.
6. Davidson, R. C., *Petroleum Refiner*, **26**, 663 (Sept., 1947).
7. Egloff, G., Morrell, J. C., Thomas, C. L., and Bloch, H. S., *J. Am. Chem. Soc.*, **61**, 3571 (1939).
8. Elkin, P. B., Shull, C. G., and Roess, L. C., *Ind. Eng. Chem.*, **37**, 327 (1945).
9. Evans, L. P., *Oil Gas J.*, **44**, 167 (March 30, 1946).
10. Good, G. M., Voge, H. H., and Greensfelder, B. S., *Ind. Eng. Chem.*, **37**, 327 (1945).
11. Greensfelder, B. S., and Voge, H. H., *Ind. Eng. Chem.*, **37**, 514 (1945).
12. ———, and ———, *Ind. Eng. Chem.*, **37**, 983 (1945).
13. ———, and ———, *Ind. Eng. Chem.*, **37**, 1038 (1945).
14. ———, ———, and Good, G. M., *Ind. Eng. Chem.*, **37**, 1168 (1945).
15. Hansford, R. C., *Ind. Eng. Chem.*, **39**, 849 (1947).
16. ———, Myers, C. G., and Sachanen, A. N., *Ind. Eng. Chem.*, **37**, 671 (1945).
17. Houdry, E. J., Burt, W. F., Pew, A. E., Jr., and Peters, W. A., Jr., *Petroleum Refiner*, **17**, 574 (Nov., 1938).
18. Krebs, R. W., *Petroleum Refiner*, **27**, 115 (May, 1948); *J. Inst. Petroleum Tech.*, **34**, A306 (1948).

19. Kutz, W. M., and Corson, B. B., *Ind. Eng. Chem.*, **38**, 761 (1946).
20. Marschner, R. F., *Ind. Eng. Chem.*, **30**, 554 (1938).
21. Matheson, G. L., Herbst, W. A., and Holt, P. H., 2nd, *Ind. Eng. Chem.*, **41**, 1099 (1949).
22. Mills, I. W., *Proc. Am. Petroleum Inst.*, **27** (3), 29 (1947); *Oil Gas J.*, **46**, 237 (Nov. 15, 1947).
23. Murphree, E. V., *Ind. Eng. Chem.*, **35**, 623 (1943).
24. ———, *Petroleum Engr.*, **19**, 86 (Sept., 1948).
25. ———, "The Fluidized Solids Technique and its Application," Paper presented at Association Francaise des Techniciens du Petrole, Paris, June 26, 1949.
26. ———, Brown, C. L., Fischer, H. G. M., Gohr, E. J., and Sweeney, W. J., *Ind Eng. Chem.*, **35**, 768 (1943).
27. ———, ———, Gohr, E. J., Jahnig, C. E., Martin, H. Z., and Tyson, C. W., *Oil Gas J.*, **43**, 64 (March 3, 1945).
28. ———, Fischer, H. G. M., Gohr, E. J., Sweeney, W. J., and Brown, C. L., *Oil Gas J.*, **42** (28), 37 (1943).
29. ———, Gohr, E. J., and Kaulakis, A. F., *Petroleum Processing*, **3**, 361 (April, 1948).
30. Nicholson, E. W., Moise, J. E., and Hardy, R. L., *Ind. Eng. Chem.*, **40**, 2033 (1948).
31. Pratt, T. W., *Proc. Am. Petroleum Inst.*, **27** (3), 38-47 (1947).
32. Ray, E. C., *Chem. Eng.*, **55**, 127-9 (May, 1948).
33. Richardson, R W., Johnson, F. B., and Robbins, L. V., Jr., *Ind. Eng. Chem.*, **41**, 1729 (1949).
34. Sachanen, A. N., "Conversion of Petroleum," 2nd ed., pp. 334-58, New York, Reinhold Publishing Corp., 1948.
35. Simpson, T. P., Eastwood, S. C., and Shimp, H. G., *Petroleum Refiner*, **24**, 110 (Nov., 1945).
36. Snuggs, J. F., *Petroleum Refiner*, **26**, 107 (March, 1947).
37. Thomas, C. L., *J. Am. Chem. Soc.*, **66**, 1586 (1944).
38. ———, Hoekstra, J., and Pinkston, J. T., *J. Am. Chem. Soc.*, **66**, 1694 (1944).
39. Van Dornick, E., *Petroleum Engr.*, **18**, 149 (April, 1947).
40. Voorhies, A. V., Jr., *Ind. Eng. Chem.*, **37**, 318 (1945).
41. Webb, G. M., *Petroleum Processing*, **2**, 497-8 (July, 1947).
42. ———, and Ehrhardt, C. H., *Petroleum Processing*, **2**, 5-8 (Jan., 1947).

B. T. Brooks, in *Science*, 111, 648-50 (June 16, 1950), suggests: ". . . hydrocarbon reactions involving carbonium ions and initiated by the catalytic action of acid silicate minerals, particularly clays, explain satisfactorily many of the problems of petroleum formation hitherto not understood."—*Ed.*

ASPHALT

R. N. Traxler

The Texas Co., Port Neches, Texas

ASPHALTIC BITUMENS are of such a complex nature that little information is available concerning the nature of the hydrocarbons present. Sachanen [1] has reviewed much of the data available on the chemical constitution of asphalts, and ultimate analyses on various fractions of different asphaltic materials were published by Hillman and Barnett.[2] All asphaltic materials contain some sulfur. It should be pointed out that ring compounds have been found to predominate in the high-boiling fractions of petroleum and that the products vary from naphthenes containing one or more rings with attached alkyl groups to hydrocarbons containing naphthene rings and an aromatic ring. The assumption that these types of compounds occur in asphaltic residua and in asphalts may be fallacious, but such an assumption gives a concept of the type of hydrocarbon present. This concept is confirmed, more or less directly, by certain facts presented in the following pages.

Logically, the physical and colloidal properties of materials as complex as asphaltic bitumens have assumed greater importance than their chemical characteristics. Empirical physical tests have been used for a long time to obtain information concerning flow characteristics, susceptibility to change in temperature, adhesiveness and other properties of practical importance, but only in recent years have certain absolute methods of measurement been developed. The realization that asphalts are colloids has been a great stimulus to the understanding of their properties and, in fact, it is only when asphaltic bitumen is studied as a colloid that real progress is made toward a comprehension of its behavior under various conditions. The idea that asphalts are colloidal systems is generally considered to have been first proposed by Nellensteyn [3] in 1924, who suggested that these materials are composed of micelles dispersed in an oily medium. He conceived that a particle of carbon formed the nucleus of each micelle and that the nucleus was surrounded by layers of adsorbed asphaltenes, each successive layer being composed of hydrocarbons of

lower molecular weight and higher hydrogen-to-carbon ratio. Dispersion of the micelles was considered to be maintained by the presence of asphaltic *resins* which were thought of as protective colloids or stabilizing agents. Nellensteyn's general theory has been widely accepted except for the idea of a carbon nucleus for each micelle.

Recently, Eilers [3a] discussed the composition and properties of asphaltenes, including comments on molecular weights by (1) cryoscopic methods, (2) measurement of viscosities of dilute solutions and (3) spreading tests on the Langmuir apparatus. Mention is made of the x-ray examination of asphaltenes by de Lange and Corbet, which followed the original work of Nellensteyn and Pfeiffer. The results of these measurements indicate the existence in the asphaltenes of six-ring planes and of paraffinic chains and, in the words of Eilers, "support the view that the asphaltenes from e.g. Mexican or Venezuelan crude oil are large molecules composed of condensed, mainly naphthenic, aromatic, hydroaromatic and thiophenic ring structures in combination with hydrocarbon chains of different lengths . . ."

Pfeiffer and his associates [4] have elaborated on Nellensteyn's theory outlined above and have proposed explanations for the flow characteristics of various kinds of asphalt on the basis of their colloidal structure. These investigators believe that hydrocarbons with the greatest molecular weight and most decided aromatic character are arranged closest to the nucleus of the micelle. Surrounding these are successive layers of lighter compounds of less aromatic nature, until a gradual and almost continuous transition into the intermicellar phase is formed. These investigators point out that there is no clear boundary between the oily constituents and resins or between the resins and asphaltenes. When the system contains sufficient *resins* or *protective colloids* to peptize fully the heavy asphaltenes, a *sol* type of asphalt results. On the other hand, a *gel* type of bitumen occurs when, owing to paucity of stabilizing agents, the micelles are not well dispersed and tend to form bonds because of mutual attractions. Pfeiffer and co-workers looked upon the gel structure as irregular open-packing, the spaces of which are filled by the intermicellar liquid.

The differences in flow and other characteristics evident in various asphalts, and discussed below, are caused by the presence of the sol or gel state or of the many intermediate gradations which may exist between the two extreme states. Much of the work done in recent years confirms the concept that asphalts are colloidal systems, and that the rheological and other properties used to evaluate a material merely reflect the degree of dispersion or solvation existing within the system. Consideration is given below to some of the most outstanding manifestations of the colloidal nature of asphaltic bitumens.

Solubilities in Organic Solvents

The high molecular-weight, dark-colored asphaltenes and resins which comprise a large portion of the micelles mentioned in the discussion above are insoluble in solvents such as light petroleum naphtha (pentane), liquid propane and normal butanol, to mention only a few; whereas, the oily maltenes or petrolenes which compose a large part of the continuous phase in the asphalt are soluble in these organic liquids. Other solvents such as acetone and furfural may be used to dissolve selectively certain of the hydrocarbons present. In every case, the amount and consequently the nature of the material dissolved, are dependent not only on the solvent used, but on its volume relative to that of the asphalt present and to the temperature maintained during the extraction.

Eilers [3a] reported extensive studies on the solubility of asphalts in a number of widely different nonpolar organic solvents. He found the per cent insoluble to correlate with the internal pressure and surface tension of the solvent. Mexican asphalt of 40 to 50 ASTM penetration at 25°C was completely soluble in nonpolar solvents with an internal pressure, $\sigma V^{-1/3}$, greater than 4.5. (In this expression σ = surface tension and V = molecular volume of the solvent.)

He also uses the viscosity of solutions of asphalt components in pure solvents as "a simple way to obtain an insight into their colloidal properties, especially so when the results are expressed in terms of the voluminosity of the solute, i.e. the volume that seems to be occupied in the solution by the quantity of solute of unit volume in the dry state." The voluminosity (v) is calculated from the Einstein equation

$$\eta r = 1 + 1.25vC$$

where ηr = relative viscosity of the solution and C = concentration of solute in parts by volume.

In Table 1, data are given for asphalts A, B and C, which illustrate the different properties associated with the sol, sol-gel and gel states respectively. At this point, attention is drawn to the fact that the amounts of high molecular-weight, dark-colored materials precipitated by the nonpolar solvents are about the same for asphalts A and C but are considerably higher for the transitional sol-gel material, asphalt B. These data indicate that the quantity of heavy materials (asphaltenes) present in an asphalt is not a major factor in determining whether the material will show sol or gel characteristics. Other more important factors will be discussed in the following pages, and asphalts A, B and C will be used to illustrate the dependence of physical behavior upon the colloidal state of the system.

TABLE 1. PROPERTIES OF ASPHALTS POSSESSING
DIFFERENT COLLOIDAL CHARACTERISTICS

Asphalt	A	B	C
Colloidal type	sol	sol-gel	gel
Method of processing		Air Blown	
Properties			
Density at 25°C	1.016	1.025	0.984
ASTM softening point, R&B°C	50	55	65.6
ASTM ductility, at 25°C, 5 cm/min, cm	200+	164	5.5
ASTM penetration, 100 gm/5 sec/25°C	50	55	53
Viscosity, megapoises* at 25°C and power input of 1000 ergs sec^{-1} cm^{-3}	3.2	4.4	23
Degree of complex flow, c	1.0	0.80	0.50
Asphalt aging index†	0.02	0.08	0.21
Elasticity, relaxation one-half time, sec	1	3	13
Solubilities			
Insoluble in pentane, %	20.6	29.7	24.2
Insoluble in liquid propane, %	37.5	44.6	38.3

* One megapoise = 1,000,000 poises
† AAI is the slope at 100 hours of a log-log plot of viscosity versus time in hours

Surface Etching by Selective Solvents

Schwarz [5] proposed a method for classifying bitumen, which comprised etching the surface of the solid bitumen with solvents such as benzene, carbon tetrachloride and carbon disulfide for a few seconds and then photographing the surface. Traxler and Coombs [6] found that ethyl ether and light petroleum naphtha were more satisfactory solvents. Definite surface patterns were evident in photographs taken at 220× with a Leitz Ultropak microscope when a gel-type asphalt was used. When a sol type was treated, little or no surface pattern was visible even though an aged asphalt surface was used. Although the obvious deduction is that some type of structure exists within the gel-type asphalt, the surface designs obtained cannot be assumed to show the structure actually existing within the asphalt.

The presence of appreciable amounts of solid paraffin wax may be detected in an asphalt by spreading a thin translucent film of the hot asphalt on a microscope slide and observing by transmitted light at 450×. If more than 1 or 2 per cent of solid paraffin wax is present, it will appear as crystal-like masses; the use of a polariscope makes the identification more positive. High solid-paraffin content, and especially overall paraffinic content are found to be associated with the gel type of asphalt. This situation may be caused in certain instances by the formation of a structure within the asphalt by the solid paraffins but more generally the gel state exists because the asphaltenes or micelles are flocculated by paraffinic hydrocarbons in the continuous phase.

Rheological Characteristics

Flow properties (both viscous and elastic deformations) at service temperatures are of primary importance in the utilization of asphaltic materials. Various empirical tests, such as the ASTM penetration tests, softening point and ductility tests, have been used by producer and consumer for many years in an attempt to evaluate these characteristics, and to determine the applicability of the asphalt for specific uses.

Figure 1. High consistency rotational viscometer showing rotor, stator, and lid.
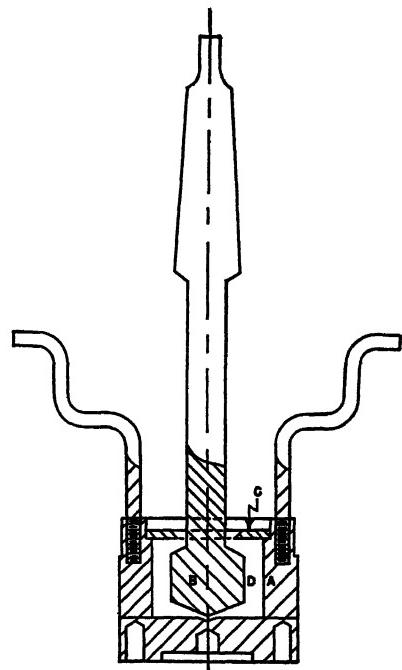
The evaluation of the rheological properties of asphalt at service temperatures in absolute (cgs) units presents many problems because of the extremely high consistencies encountered. Fortunately, rapid strides have been made during the past 15 years in the development of·apparatus capable of measuring the consistency in absolute units of materials of high consistency. The rotating cylinder type of viscometer, one form of which is described below, has proved to be very useful for this purpose. The viscometer proper is shown in Figure 1 and the entire apparatus including bath, gear housing, switches, recording device, etc., is illustrated in Figure 2.

The space, D, in Figure 1 is filled with the molten material to be tested. In order that the inner cylinder will be concentric with the outer cup, the entire assembly is held in a jig, Q, shown in Figure 2. When the

material in the viscometer has cooled to room temperature, the entire assembly is placed in the water bath and brought to the desired temperature (usually 25°C). The viscometer is connected to the driving and recording devices after constant temperature has been attained. By means of a constant-speed motor operating through a gear train, the outer cylinder, *A*, which is mounted on a turntable, is rotated at a constant, selected angular velocity (ten velocities between 0.0696 and 2200 revolutions per hour are available in the machine shown in Figure 2) and the torque required to prevent the inner cylinder, *B*, from turning is measured by a Statham gage which is enclosed in the head of the viscom-
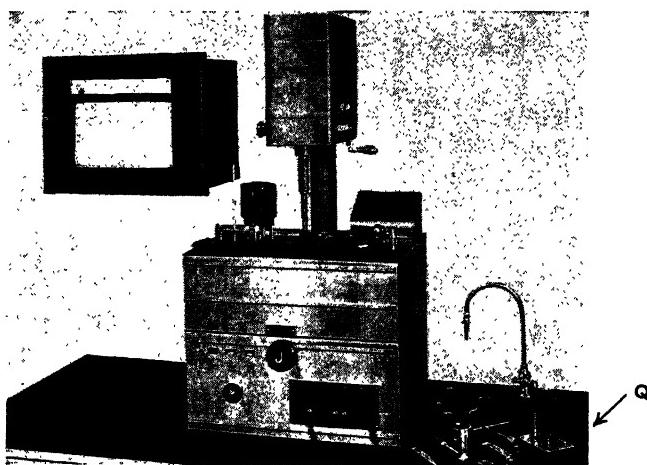
Figure 2. High consistency rotational viscometer.

eter (*H* of Figure 2). The torque is recorded on a strip chart potentiometer. A bath is provided in which the temperature variation does not exceed ±0.05°C. The top and bottom of the stator (inner cylinder) are cones of such an angle that the mean rate of shear is essentially the same at the ends as at the cylindrical part of the annulus. Lid, *C*, helps prevent elastic materials from pulling away from the stator under the shearing action applied. Viscosities are usually measured at two or three rates of shear in order to determine whether the asphalt is an essentially viscous (Newtonian) liquid or a non-Newtonian material.

A simple rheological diagram is an arithmetic plot of shearing stress in dynes cm^{-2} as abscissa versus rate of shear in sec^{-1} as ordinate. On such a plot viscous materials give straight lines, whereas non-Newtonian liquids are represented by curved lines. If the same data are plotted on log-log coordinates, a more informative rheological diagram results, since straight lines are obtained with most asphalts for both types of flow.

On the log-log type of plot, which is illustrated in Figure 3, the slope (with respect to the rate of shear axis) of the line for simple viscous flow is always 1.0, whereas that for complex flow is usually less than 1.0. Asphalt *A*, which is indicated to be an essentially simple liquid in Figure 3, is the sol-type asphalt shown in Table 1. It will be seen from the rheological diagram that asphalts *B* and *C* are represented by lines with slopes of 0.80 and 0.50 respectively. These materials are the sol-gel and gel-type asphalts described in Table 1 and discussed above. It is
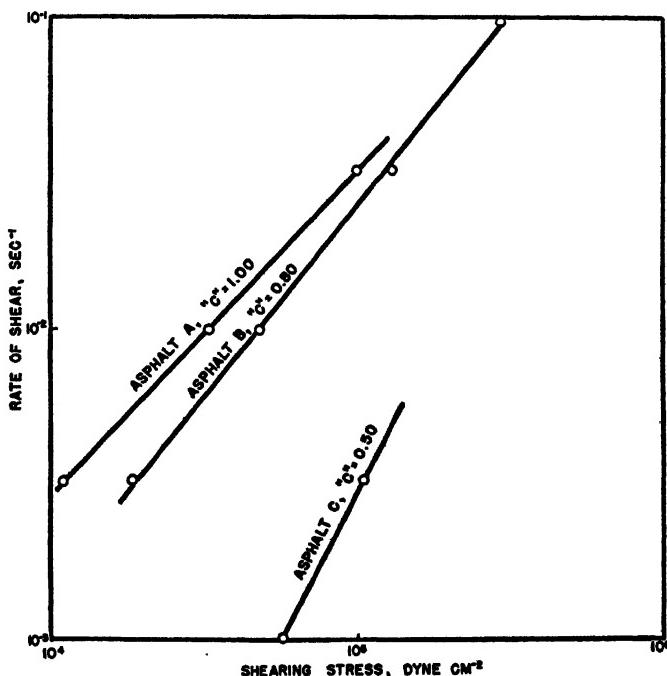
Figure 3. Rheological diagrams on log-log coordinates.

convenient to use the slope of the line on the log-log rheological diagram as a measure of the degree of deviation from viscous flow, and this value, which is listed in Table 1 as degree of complex flow, *c*, indicates the colloidal nature of the asphalt. Only the well-dispersed sol-type bitumens (e.g., asphalt *A*) show essentially viscous flow at service (atmospheric) temperatures. As the dispersion becomes less complete and gel characteristics begin to appear, the flow becomes non-Newtonian or complex in character. The value of *c* (the slope of the lines in Figure 3) may drop to 0.30 but such a low value is exceptional.

The chemical nature of the hydrocarbons in the petroleum from which an asphalt is prepared obviously affects the colloidal properties, and

consequently, the flow characteristics of the finished asphalt. Methods and degree of processing also have a marked effect on the colloidal and rheological properties of the material. Air blowing, in general, results in the appearance of more gel-type characteristics (complex flow) than are obtained by vacuum or steam distillation. The air-blowing process tends to convert the protective colloids or stabilizing agents present in the asphaltic residuum into asphaltenes, thus reducing the dispersing power of the continuous phase in the asphalt and increasing the amount of material to be dispersed. As the temperature of an asphaltic bitumen is raised, complex flow, if present, diminishes and may disappear entirely if the temperature is increased sufficiently. Illustrations of the effects of source, method of processing, degree of processing, and temperature on the rheological and, thus, on the colloidal properties of asphalt have been given by Traxler, Schweyer, and Romberg.[7]

All asphalts show some increase in consistency with time, that is not caused by loss of volatile components. Considerable experimental work involving the use of absolute viscosity data has established that the age-hardening of asphalt is dependent upon the source, the method and degree of processing, and the temperature at which the bitumen is maintained during ageing. A sol-type asphalt such as asphalt A will show a very low rate of age hardening, as indicated by the asphalt ageing index (see AAI in Table 1). As the degree of complex flow increases (value of c becomes less), that is as the asphalt becomes more definitely gel-like in nature, the rate of age-hardening (AAI) increases rapidly. This is illustrated by the data given for asphalts B and C in Table 1.

The appearance of thixotropy in asphalts is further evidence of their colloidal nature. Complex flow, age-hardening and thixotropy are all dependent upon structure within the colloidal system. As the micelles are better dispersed in the continuous phase, all three phenomena tend to decrease in magnitude.

Many of the uses for asphalt depend upon the ability of the material to deform without rupture and to recover elastically from small deformations. For viscoelastic materials such as asphalt, this recovery is partial and frequently takes place over a considerable period of time. Also, the complications encountered rule out many of the methods commonly used for determining the modulus of elasticity. Relaxation of stress has been studied recently[8] as a means of evaluating the elastic properties of asphalt. The viscometer illustrated in Figures 1 and 2 is suitable for making such measurements since it is equipped with a brake, J, for stopping the rotation quickly. The relaxation of stress is recorded on the chart, and if the relaxation is not too rapid, a fair idea of the elasticity is obtained by taking the time required for the stress to decrease

to one-half of its original value. An inspection of the data in Table 1 shows that as the degree of complex flow increases (value of c becomes smaller) the asphalt becomes more elastic. This correlation is not surprising because both phenomena probably have their origin in the same colloidal structures.

Change in consistency with temperature, commonly referred to as temperature susceptibility, is associated with the colloidal nature of the bitumen. With the degree of dispersion within the asphalt changing with temperature it is obvious that the same material may present an entirely different colloidal condition at two different temperatures. Further, the measurement of temperature susceptibility is complicated and confused by the presence of complex flow. Because of these facts most of the published data give erroneous ideas concerning the reactions of the asphalts to temperature change. Data obtained at high temperature are inadequate for estimating the properties of the same sample at atmospheric temperatures.

Dispersion and Flocculation of Asphaltenes

Mack,[9] in a pioneering study of the constitution of asphalts, investigated the dispersing power of various amounts of petrolenes from different sources for asphaltenes from several asphalts. Among his conclusions he stated: "Sols of asphaltenes from different sources in petrolenes whose oily constituents are solvents for asphaltenes, are viscous liquids of high viscosity. Asphaltene sols in petrolenes whose oily constituents do not dissolve asphaltenes, show plastic flow." In a somewhat later publication, Pfeiffer and van Doormaal [10] came to the same general conclusion and also pointed out that the aromatic nature or content of the petrolenes is responsible for any ability they may possess for dispersing the asphaltenes. When asphaltenes from a particular source are blended with petrolenes from different asphalts the mixtures have properties similar to the asphalts from which the oily fractions were obtained. For example, a blend made using petrolenes from a sol asphalt also has the properties of a sol-type material. Thus, it may be concluded that the chemical nature of the oily, dispersing (continuous) phase in an asphalt regulates to a great extent the colloidal and rheological properties of the bitumen. The asphaltenes from different sources appear to behave about the same, although obviously the amount of these high molecular-weight materials present will affect the fundamental characteristics of the asphalt.

An illustration of the dispersing and flocculating ability of the different materials present in the oily (continuous) phase of an asphalt is given in Table 2. A hard asphalt of 20 ASTM penetration at 25°C

was fractionated by treatment with *n*-butanol and the insoluble material freed of solvent. After removing the solvent from the *n*-butanol soluble portion, the resulting heavy oil was fractionated with acetone at a low temperature to dissolve out the aromatic materials of high density and

TABLE 2. DISPERSION AND FLOCCULATION OF ASPHALTENES

	Blend I	Blend II
Composition		
Asphaltenes + resins, %	42.7	28.6
Aromatic hydrocarbons, %	57.3	
Paraffinic hydrocarbons, %		71.4
Tests		
Density, at 25°C	1.048	0.994
ASTM softening point, R&B°C	52.8	53.9
ASTM ductility, at 25°C, 5 cm/min, cm	104	16
ASTM penetration, 100 gm/5 sec/25°C	77	110
Viscosity, megapoises at 25°C and power input of 1000 ergs sec^{-1} cm^{-3}	2.0	2.1
Degree of complex flow, c	0.80	0.50
Stain index	1	10

high refractive index. The material insoluble in acetone was the more paraffinic hydrocarbons of lower density and refractive index. Blends of the hard fraction (*n*-butanol insoluble) were made with each of the oil fractions to give products of about 2,000,000 poises viscosity at 25°C. It will be noted from the data given in Table 2 that blend I, made with the high density, high refractive-index material, is a transitional or sol-gel type of asphalt possessing only moderate complex flow, whereas blend II is a gel-type material with a high degree of complex flow, although it contains considerably less of the high molecular-weight, asphaltene-resin material.

Eilers [3a] has given an excellent discussion of gel structure in asphalt and uses variable colloidal structure to explain the important property of incompatibility which occurs between certain asphalts.

Stain Caused by Oil Exudation

Considerable amounts of asphalt are used as an adhesive for various purposes and especially for bonding together sheets of paper in the manufacture of shipping cartons. Certain asphalts when used for this purpose yield an oily exudate which stains the paper. Because of its unsightly appearance, such staining is considered very undesirable.

A test has been developed [11] which makes possible a quantitative comparison of the staining tendencies of different asphalts. The molten bitumen is poured into a mold, and when the material has solidified a number of discs of cigarette paper are placed in contact with the asphalt. The mold is screwed into a holder and the paper discs are held in firm

contact with the asphalt by a constant air pressure. After an established number of hours, at a constant temperature, the apparatus is taken apart and the number of stained papers are counted, omitting the sheet next to the asphalt surface. The number of stained papers is called the stain index of the bitumen. This test is not only of practical value but is of help in understanding the colloidal constitution of the asphalt. Table 3 shows the effect of the colloidal nature of four asphalts on their staining propensity. The well dispersed sol-type asphalts A and AA, which were

TABLE 3. RELATION BETWEEN STAINING PROPENSITY AND
OTHER PROPERTIES OF ASPHALT

Process	Sol-Type Asphalts		Gel-Type Asphalts	
Asphalt	A	AA	C	CC
Tests				
ASTM softening point R&B°C	50	58.9	65.6	138.9
ASTM penetration, 100 gm/5 sec/25°C	50	25	53	18
Viscosity, megapoises at 25°C and power input of 1000 ergs sec^{-1} cm^{-3}	3.2	24	23	1360
Degree of complex flow, c	1.00	0.95	0.50	0.40
Stain index	3	4	10	26

made from the same crude source and which show little or no deviation from simple, viscous flow, have low stain indices. Sample AA which had been processed (air blown) more than sample A shows slightly more staining propensity. This would be expected because the hard sample (AA) also showed slight evidence of complex flow (value of $c = 0.95$), indicating the appearance of some gel structure. On the other hand, the flocculated or less dispersed gel-type asphalts C and CC, which were both air-blown from the same residuum, show marked deviations from viscous flow and relatively high stain indices. Again, the more highly processed sample (CC) shows the greatest degree of complex flow, the most pronounced gel characteristics and the highest stain index. These examples illustrate quite clearly that staining, which results from exudation of oil from the asphalt, is a manifestation of the colloidal condition of the asphalt.

Emulsification of Asphalt

Considerable quantities of bitumen are utilized in the form of emulsions of the oil-in-water type in order to avoid using molten asphalt for certain field applications. An extensive patent literature exists,[12] covering numerous methods of emulsification and a legion of emulsifying and stabilizing agents for the preparation of emulsions for many different purposes.

An alkaline aqueous solution of low surface tension is commonly used as the dispersing phase. A detergent material (e.g. soap) is especially effective as the solute, because the soap molecules concentrate at the bitumen-aqueous interface in such a manner that the hydrocarbon part of the soap is directed toward the asphalt surface and the carboxyl end of the molecule toward the water phase (Langmuir-Harkins theory). The effect of any protective colloids adsorbed on the asphalt surface is to prevent the too ready coalescence of the droplets of bitumen. Dispersions of bentonite clays have been used for many years as emulsifiers and stabilizers, and the so-called clay emulsions possess certain distinctive characteristics such as the drying of a film from the bottom up instead of from the top down, as in the case of the soap-stabilized emulsion.

Most commercial emulsions are made in some kind of a colloid mill, although certain asphalts are so readily emulsified that the only equipment required is a good stirrer. Bitumens having a high aromatic hydrocarbon content and possessing the sol characterization discussed above are usually rather easily emulsified. When gel properties begin to appear, emulsification becomes more difficult. The difficulties also appear to be accentuated by a low acid number for the asphalt. It cannot be said that a low (less than 1.0) acid number always indicates poor emulsifiability, but an asphalt with a high acid number (more than 1.0) is seldom found to be unpromising as an emulsion base.

In the use of either a simple mixer or an elaborate colloid mill for the manufacture of emulsion, the molten asphalt is brought into contact with the hot aqueous solution or dispersion. Most colloid mills create a thin film of the bitumen which in the presence of the aqueous phase breaks up into small droplets; in some dispersing machines thin threads of bitumen are formed which also divide to yield the droplets which usually are 1 to 5 microns in diameter. The smaller the particle size the more stable the emulsion is likely to be; also particle size and size distribution are important factors in determining the consistency and flow properties of the emulsion. Two emulsions of the same phase composition may have different viscosities because of differences in particle size of the dispersed bitumen. Such differences may be obtained by varying the temperature of the asphalt and water and the shearing stress applied in preparation of the emulsion. It has been demonstrated [13] that particle size distribution has a pronounced effect on emulsion consistency. A comparison of emulsions having the same phase composition and same volume per cent of dispersed phase showed that nonuniformity in particle size of the dispersed phase resulted in emulsions of lower viscosity. This principle is illustrated in Figure 4, which was taken from the publication referred to above. Emulsion *A* contained particles of uniform size which were less than 1 micron in diameter; emulsion *B* also was made up of

uniform particles but they were larger (2 to 3 microns in diameter). A decided drop in viscosity occurred when the two emulsions, similar in every way except for particle size, were blended. Creation of a wider range in particle size distribution by other means also results in decreased viscosity. It must not be assumed, however, that particle size and size distribution are the only factors influencing the rheological properties of asphalt emulsions. The nature and concentration of the colloidal materials present in the continuous phase, the presence of electrolytes, the pH of the medium, and probably other variables influence the properties of the emulsion. For example, an increase in the amount of a
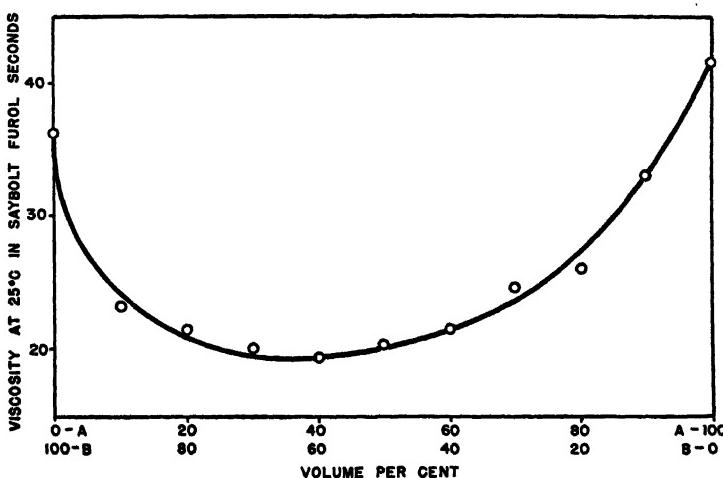
Figure 4. Effect of particle size distribution on viscosity of asphalt emulsions.

protein stabilizer, such as casein, soybean flour or cottonseed meal, results in an increase in emulsion consistency for a given asphalt content and particle size distribution.

Lyttleton and Traxler [14] investigated the flow properties of over thirty different emulsions and concluded that each emulsion belonged to one of three groups, namely, plastic flow, inverted plastic flow or Newtonian flow, in that order of frequency. A commercial asphalt emulsion possessing Newtonian flow is a rarity; most of them are non-Newtonian liquids which show a decrease of consistency with increasing rate of shear. The non-Newtonian type which shows an increase in consistency with increasing rate of shear (the inverted plasticity type) is not usually encountered.

If a series of asphalts made from a particular source are emulsified with the same emulsifying and stabilizing agents, the hardest asphalt is the most difficult to disperse. Also, it is usually difficult to obtain an

emulsion of proper stability when an asphalt of high consistency is employed.

The involved forces which govern the stabilities of asphalt emulsions under various conditions present a challenging problem in colloid chemistry that no one apparently has attacked with sufficient vigor and ingenuity. The importance of the nature of the highly adsorbed lyosphere around the droplet of asphalt, which is composed chiefly of bound water and emulsifier, and the effect of electrolytes present in altering the charge on the particles and as dehydrating agents, are worthy of careful investigation. Effects of bacterial action on asphalt emulsions, especially on those containing protein stabilizers should not be ignored. The products of bacterial decomposition of the stabilizing agent, which result in changes in the pH and composition of the aqueous phase, may be responsible for the changes in stability and consistency sometimes noted in emulsions subjected to several months of storage.

Uses for Asphalt

Approximately 60 per cent of the asphalt manufactured is consumed in the construction of roads and airports. Since there are numerous kinds of roads and pavements designed for various climatic and service conditions, it is to be expected that the properties of the bitumen used will vary considerably. Asphalt is employed chiefly in road building as a binder for the mineral aggregate (crushed stone, etc.); it also serves to prevent the entrance of water into the road surface. In certain types of construction, molten asphalt and hot stone are thoroughly mixed, laid on the prepared road bed, and then compacted by means of a heavy roller. For building rural roads and for situations where the use of hot asphalt would be inconvenient, the asphalt frequently is reduced to a fluid consistency at atmospheric temperatures by means of various grades of naphtha and the cutback asphalt mixed cold with the aggregate or soil. In this type of road construction most of the solvent is allowed to evaporate before traffic is permitted on the road. A sol-gel type of asphalt is commonly used in road building and the viscosity usually does not exceed 20 million poises at 25°C. The consistency desired depends to a great extent upon the climatic conditions and atmospheric temperatures to which the road will be subjected. Good adhesion to the aggregate is an important prerequisite for a satisfactory road-building asphalt.

Asphalt emulsions have, to some extent, supplanted cut back asphalts in the construction of roads, footpaths, runways, etc., where the use of hot asphalt would be impractical. The emulsion used for any particular purpose is generally selected on the basis of its stability—that is, the rapidity with which the asphalt "breaks out" on the soil, sand or stone.

The breaking of an asphalt emulsion under such conditions is caused by the evaporation of the water and the inactivation of the stabilizing agents or protective colloids by small amounts of soluble salts picked up from the surface of the aggregate. In countries where the supply of petroleum is limited, emulsions are widely used in road building and other large projects in order to conserve the naphtha which would be consumed in the manufacture of cutbacks. Asphalts for the preparation of emulsions are usually of the sol type, with viscosities of 100,000 to 1,000,000 or more poises at 25°C.

A large volume of asphalt is used in the manufacture of prepared roofing for homes and stores and in the preparation of built-up roofs for industrial buildings. In the manufacture of prepared roofings, a sheet of porous paper, called roofing felt, is dipped into hot asphalt, which drives out the air and moisture and fills the voids in the paper with the asphalt. The saturated sheet is then coated on both sides, at a high temperature, with a hard asphalt which is selected for its resistance to deterioration by the weather. The asphalts selected for saturating the felt usually vary between one and 30 million poises at 25°C, whereas the viscosity of the coating asphalts may exceed 500 million poises at that temperature. Gel-type asphalts are used in roofing manufacture and of course, as has been pointed out above, the harder the asphalt the more pronounced are the gel characteristics. Some types of prepared roofings are dusted with fine mineral powder to prevent sticking of the asphalt coating when the roofing is rolled for shipment and storage. The better-prepared roofings have carefully graded colored granules pressed into the top surface during manufacture. These colored granules lend beauty to the roof and protect the asphalt surface from the action of sunlight, air, and rain, resulting in a longer service life.

In the preparation of build-up roofs, used extensively on industrial buildings, successive layers of asphalt-saturated roofing felt and hot asphalt of the proper consistency are applied to the roof deck. The top layer of asphalt frequently has gravel or other mineral aggregate embedded in it to offer protection against the weather and any foot traffic that may occur.

Numerous other uses and special applications for asphalt have been developed, which are commercially important although they consume a relatively small percentage of the total amount of asphalt manufactured. Paints, insulation, sound-deadening, waterproofing, rust-prevention, and paper-plying materials are examples of the special products encountered. Particular combinations of physical, rheological, and colloidal properties are necessary for these special purposes. For example, in paper plying, an asphalt possessing good adhesiveness for the paper and low stain-

ing propensity combined with the proper rheological characteristics is required.

The many uses that have been developed for asphalt depend primarily upon its colloidal properties, which are reflected in its unique and varied rheological characteristics, adhesiveness for solid surfaces, water-proofing properties, and resistance to deterioration by the weather.

REFERENCES

1. Sachanen, A. N., "The Chemical Constitution of Petroleum," New York, Reinhold Publishing Corp., 1945.
2. Hillman, E. S., and Barnett, B., *Proc. Am. Soc. Testing Materials,* 37 (II), 558 (1937).
3. Nellensteyn, F. J., *J. Inst. Petroleum Tech.,* 10, 311 (1924).
3a. Eilers, H., *J. Phys. Coll. Chem.,* 53, 1195 (1949).
4. Pfeiffer, J. P., and Saal, R. N. J., *J. Phys. Chem.,* 44, 139 (1940).
5. Schwarz, F., *Asphalt und Strassenbau,* 5, Supplement of *Petroleum Z.,* 28 (1932).
6. Traxler, R. N., and Coombs, C. E., *Ind. Eng. Chem.,* 30, 440 (1938).
7. ———, Schweyer, H. E. and Romberg, J. W., *Ind. Eng. Chem.,* 36, 823 (1944).
8. Romberg, J. W., and Traxler, R. N., *J. Colloid Sci.,* 2, 33 (1947).
9. Mack, C., *Assoc. Asphalt Paving Tech., Proc. Tech Sessions,* 40 (Dec., 1933).
10. Pfeiffer, J. P., and van Doormaal, P. M., *J. Inst. Petroleum Tech.,* 22, 414 (1936).
11. Knowles, E. C., McCoy, F. C., Schweyer, H. E., and Wilkinson, C. E., "Study of Oil Exudation Property of Asphalts," presented at Meeting of Am. Chem. Soc., San Francisco, Calif., March 27-April 1, 1949.
12. Abraham, H., "Asphalts and Allied Substances," 5th ed., Chap. XXV, pp. 1653-78, D. Van Nostrand Co., Inc., 1945.
13. Traxler, R. N., *Ind. Eng. Chem.,* 28, 1011 (1936).
14. Lyttleton, D. V., and Traxler, R. N., *Ind. Eng. Chem.,* 40, 2115 (1948).

APPLICATIONS OF ANTONOFF'S LAW

F. J. Nellensteyn

Rijkswegenbouwlaboratorium, The Hague, Netherlands

Surface-Tension Rule

The surface-tension rule [1] gives the relationship between the surface tension of liquids which are completely miscible with the medium, the protective bodies of asphaltic bitumen, and the flocculating or peptizing properties of the mixture. When the surface tension of such liquids exceeds 26 dynes/cm at 25°C, asphaltic bitumen is completely peptized. Liquids having a surface tension below 24 dynes/cm at 25°C are flocculents. Between 24 and 26 dynes/cm an intermediate zone is observed wherein most asphaltic bitumens are dissolved, but some of the less stable systems are flocculated. Here, however, the surface tension does not act directly, but rather the interfacial tension micelle/medium, which Antonoff's law relates to surface tension by the following equation:

$$\sigma_{micelle/medium} = \sigma_{micelle/air} - \sigma_{medium/air}$$

When $\sigma_{medium/air}$ is decreased by addition of a liquid of low surface tension, the $\sigma_{micelle/medium}$ is increased; and this causes flocculation below a certain critical value of the solvent.

The surface-tension rule has been tested out with many liquids of quite different natures; aliphatic and aromatic organic, as well as inorganic compounds. Although the interfacial tension$_{micelle/medium}$ is not the only factor controlling these phenomena, the single exception found certainly may not be regarded as an exception to Antonoff's law.

No sharp quantitative relationship exists between the amount of flocculate and the surface tension of the flocculant, although the general tendency is for liquids of higher surface tension to yield smaller amounts of flocculate. In the case of tar, however, Duriez [2] deduced a formula based on the same principles, which gives a rather sharp relation between surface tension and flocculate:

$$\log \frac{k}{p - p_0} = a\sigma^2$$

where σ represents the surface tension, p the amount of flocculate, and p_0, k, and a are parameters.

Application of Antonoff's Law to Liquid-Liquid-Liquid and Solid-Gas-Liquid Systems

For systems consisting of three immiscible liquids, l_1, l_2, and l_3, the usual formulation of Antonoff's law must be modified as follows: The three interfacial tensions are $\sigma_{l_1 l_2}$, $\sigma_{l_1 l_3}$, and $\sigma_{l_2 l_3}$. If the first of these represents the highest value, then

$$\sigma_{l_1 l_2} = \sigma_{l_1 l_3} + \sigma_{l_2 l_3}$$

Antonoff's law for the liquid-liquid-gas system may be similarly written

$$\sigma_{l_1 g} = \sigma_{l_2 g} + \sigma_{l_1 l_2}$$

where liquid l_1 shows the highest surface tension.

This generalization of Antonoff's law, indicating that in a three-phase system one interfacial tension is always equal to the sum of the two other interfacial tensions, leads to important conclusions for the solid-liquid-gas system.[3] Here we get three equations:

$$\sigma_{sg} = \sigma_{lg} + \sigma_{sl} \tag{1}$$

$$\sigma_{lg} = \sigma_{sg} + \sigma_{sl} \tag{2}$$

$$\sigma_{sl} = \sigma_{sg} + \sigma_{lg} \tag{3}$$

These equations represent three possible wetting conditions: (1) for complete wetting; (2) for incomplete wetting; and (3) for complete nonwetting. This last condition cannot be realized in the solid-liquid-gas system, but only in the solid-liquid-liquid system.

The surface tension is in reality free surface energy;[4] the surface-tension equations may also be deduced as a free surface-energy problem. Keeping this in mind, the validity of Antonoff's law may be easily checked for the solid-liquid-gas system, in the case of complete wetting.

The rise of liquid in a nonwetted capillary tube is equal to that in a capillary tube which has previously been wetted.[5] In the first case the rise is caused by $\sigma_{sg} - \sigma_{sl}$, in the second by σ_{lg}. Therefore

$$\sigma_{sg} - \sigma_{sl} = \sigma_{lg} \quad \text{or} \quad \sigma_{sg} = \sigma_{lg} + \sigma_{sl}$$

Incomplete wetting may develop an acute, a right, or an obtuse contact angle.

Determination of the Free Surface Energy of Solids

In the case of an acute contact angle, e.g., in the system glass-mercury-air, the free surface energy of the glass may be calculated from the fall

of the mercury in the glass capillary, the adhesion tension $a = \sigma_{sl} - \sigma_{sg}$.
In the case of incomplete wetting,[6]

$$\sigma_{lg} = \sigma_{sl} + \sigma_{sg}$$

Whence

$$\sigma_{sg} = \tfrac{1}{2}(\sigma_{lg} - a)$$

A variation of this method has been given by Loman and Zwikker,[7] who calculated a from the angle of contact of mercury on various minerals.

REFERENCES

1. Nellensteyn, F. J., in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. III, p. 537, New York, Reinhold Publishing Corp., 1931; Nellensteyn, F. J., and Loman, R., Permanent International Association of Road Congresses, Washington, D. C., 1930, Report 2-C.
2. Duriez, M., *Ann. Ponts et Chaussées,* 108, 403 (1938).
3. Nellensteyn, F. J., *Chem. Weekblad,* 35, 283 (1938).
4. ——, *Chem. Weekblad,* 24, 54 (1927); 35, 283 (1938); Bikerman, J. J., "Surface Chemistry for Industrial Research," pp. 1 ff., New York, 1947.
5. Nellensteyn, F. J., and Loman, R., "Asfaltbitumen en teer," pp. 46 f., Amsterdam, D. B. Centen's Uitgevers My., 1932.
6. ——, and Roodenburg, N. M., *Chem. Weekblad,* 30, 190 (1933).
7. Loman, R., and Zwikker, N. P., *Physica,* 1, 1181 (1934).
8. Nellensteyn, F. J., and Baart, M., *J. Inst. Petroleum,* 35, 302 (1949).

THE RELATIONSHIPS BETWEEN CONSTITUTION AND MODE OF ACTION OF SOAPLIKE COLLOIDAL ELECTROLYTES

A. Chwala

Technical University, Vienna
(Translated by Jerome Alexander)

THE EXTRAORDINARY GROWTH which has taken place in the production of synthetic soaplike agents, especially the so-called anion soaps,[1] makes it seem desirable to examine the extent of our knowledge of the relation between the chemical constitution of these soaplike colloidal electrolytes and their colloidal properties.

By way of anticipation it should be said that exhaustive investigations during the last twenty years in the field of soaplike colloidal electrolytes, especially on their state of aggregation in water dispersion, have led to a series of interesting particulars.[2] What may be looked upon as a fundamental result of these investigations of various anionic and cationic soaps * is the fact that above a certain concentration known as the "critical concentration," anionic and cationic soaps can exist mainly in an aggregated state, that is, in the form of so-called ionic micelles, concerning which opinions still diverge as to the mechanism involved and kind of aggregation as well as to the size of the aggregates, i.e., the micelles. A fact which is of significance in practice is that soaplike colloidal electrolytes, predominantly the anionic soaps, are always used in such concentrations that the compounds are dispersed in the water only in the form of colloidal aggregates. Hence it is clear that universal relationships between the constitutional fine-structure of anionic and cationic-soaps and their colloid-chemical behavior must always depend functionally upon the fine structure of their micelles. Since, however, direct methods for the determination of the fine structure of micelles are lacking, such a relationship between micelle structure and the constitution of capillary-active substances cannot be directly indicated. On the other hand, the capillary action of anionic and cationic soaps is

---

* Conductivity measurements are mainly involved here, as well as determinations of electrochemical nature.

TABLE 1. VARIOUS SOAPS AND MODIFIED SOAPS

Class Designation	Chemical Nature	Chemical Formula (R = aliphatic hydrocarbon residue)	Trade Name
Soaps	Alkali and amine-salts of higher molecular fatty acids	$RCOONa [K,N(C_2H_4OH)_3]$	
Modified and polyvalent soaps	Oleylsarcoside-sodium	$RCONCH_2COONa$ 丨 CH_3	Medialan A
	Alkylsulfamide derivatives	RSO_2NHCH_2COONa	Emulphor STH
	Protein—oleic acid chloride condensation products	$RCONHR'CONHR''$ 丨 $COONa$	Lamepon A
	Sodium alkylmalonate	$R{-}CH$〈$\substack{COONa \\ COONa}$	None
	Polyacrylate	$\left[\substack{-CH-CH_2-\\ COONa}\right]_x$	Rohagit
	Sodium cellulose glycolate (methoxycarbonate)	$Cell{-}O{-}CH_2COONa$	CMC Tylose Relatin

directly connected with the fine structure of the micelles, so that one can characterize such relationships without being able to give more precise data as to the structure of the aggregated anionic or cationic soap particles. In this connection, it should be mentioned that we do have more exact knowledge of the nature and structure of such aggregates which are measurable röntgenographically. Such relatively coarsely dispersed particles appear at a concentration (known as "critical concentration," and called the "röntgenographic index") which is, however, considerably higher than that "critical concentration" at which, on the basis of conductivity measurements, a sudden change in the state of aggregation must take place. While the latter, in the case of the usual commercial products, sets in at concentrations of less than 1 gram per liter, the röntgenographically detectable colloidal particles appear only at concentrations of 30 to 50 grams per liter. As is known, one works in practice with solutions which contain 1 to 2 grams per liter, so that the röntgenographically measurable colloidal particles may be neglected in considering the relationship between molecular fine structure and capillary-active phenomena.

It may furthermore be considered as proven on the basis of recent

TABLE 2. STATUS OF VARIOUS SULFURIC ACID ESTERS

Class Designation	Chemical Nature	Chemical Formula (R = aliphatic hydrocarbon residue)	Trade Name
Sulfonated oils and fats, mainly castor oil	(a) Salts of sulfuric acid, esters of oleins and fats with free carboxyl groups and intermediately placed sulfo groups	$R\text{—CH—}R'\text{—COOH(Na)}$ $\quad\quad$ | $\quad\quad OSO_3Na$	Turkey red oil Higher sulfonates
	(b) Salts of sulfuric acid esters of oleins and fats with carboxyl groups blocked (α) by low molecular alcohols, e.g., Na-sulfonate of castor oil butylate esters (β) by aliphatic or aromatic amines, e.g., Na-sulfonate of castor oil—N-methylanilid	$R\text{—CH—}R'\text{—COOC}_4H_9$ $\quad\quad$ | $\quad\quad OSO_3Na$ $R\text{—CH—}R'\text{—CON}\langle$⬡$\rangle$ $\quad/ \quad\quad\quad /$ $OSO_3Na \quad CH_3$	Avirol AH extra Humectol CX
Fatty alcohol sulfonates (alkyl sulfates O-sulfonates)	Normal, primary alkyl sulfates (obtained from natural fatty alcohols or by hydration of natural fatty acids)	$R\text{—O—}SO_3Na$	Gardinol (Duponol) Fewa
	Secondary alkyl sulfates (from synthetic fatty alcohols, e.g., from petroleum or from fatty alcohols by the "OXO" process)	$R\text{—CH—}R'$ $\quad\quad$ | $\quad\quad OSO_3Na$	"Tergitol" Teepol
Fatty acid ethanolamine sulfonates	Alkali salts of sulfuric acid esters of higher molecular fatty acid ethanol amides	$RCONHC_2H_4\text{—}OSO_3Na$	Alframin Sulframin
Sulfonates of mono-fatty acid esters	Fatty acid monoglyceride sulfate	$RCOOCH_2CHOHCH_2OSO_3Na$	Vel
of polyalcohols (glycol, glycerin, pentaerythrit)	Alkylthiosulfate	$R\text{—S—}SO_3Na$	Only of theoretical interest

work [3] that the capillary activity of anionic and cationic soaps involves the formation of an adsorption layer one or, at the most, several molecules thick. The specificity of surface activity involves an additional factor to be considered, namely the relation between detailed molecular structure and the ability to form the most perfect molecular layers.

TABLE 3. VARIOUS SULFONATES

Class Designation	Chemical Nature	Chemical Formula (R = aliphatic hydrocarbon residue)	Trade Name
True sulfonates (C-sulfonates)	Normal alkyl sulfonates	$R—SO_3H(Na)$	Only of theoretical interest
	Secondary alkyl sulfonates made from petroleum by direct sulfonation (unimportant) or by chlorosulfonation or sulfooxidation from paraffin hydrocarbons	$R—\underset{\underset{SO_3Na}{\big|}}{CH}—R'$	Mersolat H, 30 Duponol 646B
Modified alkyl sulfonates (fatty acid condensation products, etc.)	Fatty acid esters of oxyethane sulfonates	$RCOOC_2H_4SO_3Na$	Igepon T Arctic SyntexA
	Diester of sulfosuccinic acid	$NaSO_3—\underset{\underset{CH_2—COOR\,(R')}{\big|}}{CH}—COOR$	Aerosol Deceresol
	Fatty acid methyl taurides	$RCON(CH_3)C_2H_4SO_3Na$	Igepon T Arctic SyntexT
	Fatty acid amides of aminomethane sodium sulfonate	$RCONHCH_2SO_3Na$	Igepon Substitute
Aromatic sulfonates (alkyl aryl sulfonates)	Higher molecular esters of sulfobenzoic acid	COOR ... —SO₃Na (benzene ring structure)	
	Alkyl benzol sulfonates (e.g., sodium dodecyl benzoate)	R ... —SO₃Na (R generally branched)	Nacconol NR Santomerse D Oronite
	Sulfonates of higher molecular phenones (palmitophenone)	RCO⟨ ⟩—SO_3Na	Melioran F 6
	Alkyl naphthalene sulfonates	R ... —SO₃Na (R generally short-chained)	Nekal
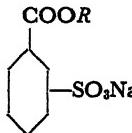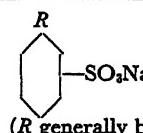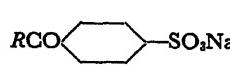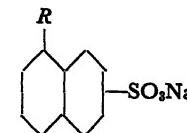
TABLE 3 VARIOUS SULFONATES—*Continued*

Class Designation	Chemical Nature	Chemical Formula (R = aliphatic hydrocarbon residue)	Trade Name
Sulfonates of hetrocyclic compounds	Alkyl benzimidazol sulfonates		Ultravon
	Alkylamide azoline sulfonates		
	Alkly triazine sulfonates		Eriopon Ac
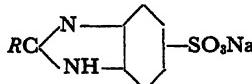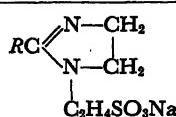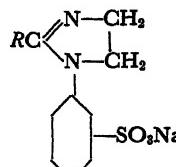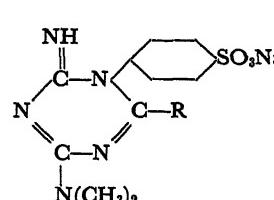
Obviously, the thicker and more firmly packed the molecules in such an interfacial layer, the greater will be its stability as compared with a layer in which single molecules are held in loose order. On the other hand, the formation of a well-ordered interfacial layer demands more time and energy than one of irregular structure. This consideration has practical significance.

Until comparatively recent times the only soaplike colloid electrolytes known (especially among anionic soaps) were for the most part products made from natural fats and oils; but in the last decade numerous new products have been developed whose molecular structure is quite different. It is thus possible to test out the properties of these modified anionic soaps and to work out in broad outline the relation existing between their molecular and micellar structure and their capillary activity. In order to get the overall picture, it is advantageous to consider, in Tables 1, 2, and 3, the anionic soaps at present known, and their chemical constitutions.

Classification of Anionic Soaps

These three tables show that anionic soaps fall into several large classes, each sub-group containing compounds whose hydrocarbon residue is either a straight chain or only slightly branched, or it shows a more or less marked branching. It is known that the specificities of the anionic and cationic soaps are in every instance consequent on a hydrophobe hydrocarbon residue of high molecular weight being brought into aqueous dispersion by an ion-active group which is solvated by water. It is merely the balance of forces between the solvation tendency and the water-repelling tendency of the hydrocarbon residue that determines the labile condition of such particles in aqueous dispersion, and alterations in this balance must show itself in variations in the capillary-active behavior. Experience has shown, especially in the study of homologous series of anionic and cationic soaps, that these changes in capillary activity are relatively small when compared in regard to the surface or interfacial tension, and in general yield no extensive insight on practical tests. Quite different theoretical considerations are suggested by measurements made on diverse substances, as shown by anionic soaps found in use recently.

For the first time there emerges the great influence of the configuration of the hydrocarbon residue on the capillary activity. It was, indeed, already known that on making sulfates from unsaturated oleyl alcohol, the interior arrangement of the sulfo acid ester groups has an influence on the effectiveness of washing, but it was not then possible to gain a general view of the connection between constitution and capillary activity. Recent work on isomer mixtures of fatty alcohol sulfates, alkylsulfonates and alkylarylsulfonates unite in showing that such a relation exists to a greater extent than might be expected, and also shows the nature and direction of such differentiation. It must be emphasized that such differences reflect only partially the influence of surface or interfacial tension, but show up in practical use by giving desirable or undesirable properties.

Before going into details as to utilization of these properties, the following general remarks are in order. Molecules having the same number of carbon atoms per molecule show, with increasing branching of the alkyl residue, a progressive decrease in length measured in the direction of the main chain. Experience indicates that it is not practical to have the length of the straight alkyl residue less than a certain minimum, and this independent of greater or lesser branching. Branching of the alkyl residue may exist in the raw material (i.e., in such products as are derived from the naturally highly branched hydrocarbon mixtures found in petroleum); or the branching may first arise because

of the process used in making the product in question, as is the case with alkane sulfonyl chlorides or the sulfonates derived from them. This opens the door somewhat to the possibility of controlling the relative amounts of branched chains and straight chains. Suitable choice of raw materials permits a certain voluntary degree of variation. In the case of processes resembling the above-mentioned chlorosulfuric acid reaction products, wherein the sulfuro-chloridization may attack any portion of the molecule, the choice of a linear starting material does not, preclude complete branching. This is evident from the following diagram,[4] applicable especially to the isomeric cetane sulfonates.

(1) $C_{16}H_{33}SO_3Na$

(2) $C_{14}H_{29}CH—CH_3$
    $\quad\quad\quad |$
    $\quad\quad\quad SO_3Na$

(3) $C_{13}H_{27}CH—C_2H_5$
    $\quad\quad\quad |$
    $\quad\quad\quad SO_3Na$

(4) $C_{12}H_{25}CH—C_3H_7$
    $\quad\quad\quad |$
    $\quad\quad\quad SO_3Na$

(5) $C_{11}H_{23}CH—C_4H_9$
    $\quad\quad\quad |$
    $\quad\quad\quad SO_3Na$

(6) $C_{10}H_{21}CH—C_5H_{11}$
    $\quad\quad\quad |$
    $\quad\quad\quad SO_3Na$

(7) $C_9H_{19}CH—C_6H_{13}$
    $\quad\quad\quad |$
    $\quad\quad\quad SO_3Na$

(8) $C_8H_{17}CH—C_7H_{15}$
    $\quad\quad\quad |$
    $\quad\quad\quad SO_3Na$
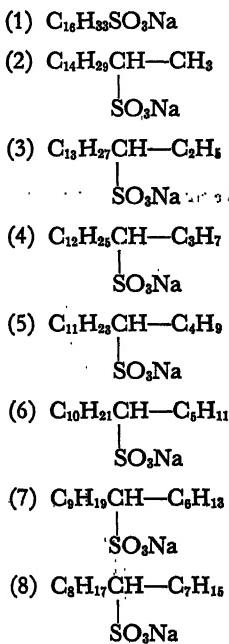
In this chosen instance the branching cannot be arbitrarily influenced. At best, by starting with a hydrocarbon of selected chain length, the shorter alkyl residues formed by treatment with sulfur dioxide and chlorine may be somewhat different in length; but as the diagram indicates, branching once established cannot then be altered.

These considerations lead to the view that with modern anionic soaps we are often dealing not with individual chemicals as is the case with soaps, with the original fatty alcohol sulfonates, or with the so-called fatty acid condensation products; we are dealing rather with a mixture of various isomers which are formed even if we start out originally with a single individual substance, cetane for example. This last assumption, however, does not apply in actual practice, where

the raw material itself is a mixture of various hydrocarbons with branching forms and different chain lengths. It is evident, therefore, that under such conditions the final product must consist of mixtures of different isomeric sulfonates having quite different chain lengths. In such an indefinite mixture it is obviously impossible to determine or to describe with scientific accuracy the relationship between molecular constitution and capillary activity, because the diversity of the chemical individuals leads to a criss-crossing of properties, and only an average resultant is determinable. However, by selecting certain typical substances made from a single substance which could be regarded as a chemical individual, it was possible by roundabout methods to produce products which though containing isomers, nevertheless had exactly the same number of carbon atoms per molecule. Investigations with such substances could for the first time lead to generally applicable conclusions.

Principles Governing the Behavior of Anionic Soaps

As a result of such experiments made by various investigators[5] the following general principles emerged:

(1) For a homologous series of anionic soaps the capillary activity at a given temperature shows an optimum value where there is a certain relation between the hydrophobe and the hydrophile portion of the molecule. Obviously this is tied to an optimum lability of the colloidally dispersed particles. Of course the degree of dispersity can be influenced by the addition of electrolytes to their water solution, over and above the relationship between the hydrophobe and hydrophile parts of the molecule of the soap-like substance. Particles too finely dispersed, e.g., with small or highly branched hydrocarbon residues, have their capillary-active properties improved by slight additions of electrolytes. Capillary-active substances with intermediate length of hydrocarbon residue are only slightly improved by addition of electrolytes. In all cases excessive additions of foreign electrolytes lead to reduction of interfacial activity.

(2) On the introduction of a second or a third hydrophilic electrovalent group the capillary activity is reduced or even increased, in which case it is immaterial whether the hydrophile electrovalent groups adjoin or are separated by hydrophobe hydrocarbon residues.

(3) The capillary activity, as determined by wetting and permeability tests, is favorably affected by very extensive branching of the hydrocarbon residue, providing the straight chain unbranched portion does not fall below a certain minimum.

On the basis of these general considerations the following conclusions

may be drawn as to the use of anionic soaps, conclusions which are continually confirmed in actual practice:

(a) Interfacial behavior is certainly determinative as a practical test for the various anionic soaps, but not always exclusively so. Other colloid chemical properties, e.g., colloid protection, may also exercise a marked influence.

(b) The wetting-, dispersing-, and cleansing agents in practical use appear to have only one electrovalent group in the molecule.

(c) With powerful wetting agents, used mainly to wet and penetrate textile materials, without necessarily developing marked dispersive power and the like, it is in general necessary that the alkyl residue (or residues) of the molecule be as widely spread out as is possible—always assuming that there is proper balance between the hydrophobe and the hydrophile part of the basic molecule. Here is its relatively unimportant whether the solubilizing electrovalent groups are directly attached to the alkyl residue, or whether the alkyl residue is carried by an aromatic or hydroaromatic hydrocarbon skeleton, or else another type of atom (e.g., oxygen of a COO—group) is inserted between the hydrophile groups. The position of the electrovalent, anion-active groups in a molecule has no pronounced influence on the wetting power.

(d) In contradistinction to what is stated above under (c) regarding wetting power, real washing and dispersion agents require for optimum activity as straight and as unbranched an alkyl residue as possible. The nearer the solubilizing anion-active group is to the end, the better the washing power. It has, however, been definitely established that higher molecular soap-like substances (anionic soaps) develop a very considerable washing power, if their alkyl residue is more or less branched, or if the anion-active group is placed more or less centrally. In the latter case it is only necessary to make sure that the branched alkyl residue does not exceed minimum dimensions. The nature of the anion-active group is not critical here. The same rules govern with COONa-, SO_4Na-, and SO_3Na- groups, i.e., with true soaps, with fatty alcohol sulfates, and, with real alkyl sulfonates or alkyl aryl sulfonates.

The fine structure of the hydrophobic part of the molecule, i.e., the position of the anion-active, electrovalent group, is also of great significance, because other properties in addition to interfacial activity are thereby affected. The washing process demands more of the washing agent than mere wetting and dispersion of the particles of soil. There must also be adsorptive power and the ability to form a stable adsorption layer around the dispersed particles of dirt. These are properties known to the practical man as "dirt-carrying power" or "washing reserve." For example, if the colloid nature of the washing material is depressed by too great a branching of the hydrocarbon residue, the dirt-

carrying power is diminished and the washed goods become "gray," because the already dispersed soil particles are again readsorbed by the textile fibers. In cases where a natural tendency to marked branching cannot be overcome by deliberately influencing the hydrophile and the hydrophobe portions of the molecule, say by choice of an initial substance of higher molecular weight, it is nevertheless possible to secure a practically satisfactory washing agent by adding other aidful substances having marked protective colloidal action, e.g., sodium cellulose glycolate.

(e) In addition to wetting and washing capacity, the colloid chemical properties are quite generally depressed by increasing branching of the hydrophobe hydrocarbon residue in the direction of increasing hydrophilicity. In such cases where the main valence chain contains 10 to 14 carbon atoms, the interfacial activity is in general favorably influenced. If the basic molecule has 14 to 18 carbon atoms, marked branching of the hydrophobic hydrocarbon residue is advantageous, because it leads to clear, concentrated solutions. The main disadvantage of a too highly branched alkyl residue, or an anion-active group in too central a position, lies in the fact that the hygroscopicity of such anionic soaps is greatly increased, making it difficult or impossible to produce the product in powdered form.

REFERENCES

1. McCutcheon, *Chem. Ind.,* 811 (1947). See also Sittenfield, *Chem. Eng.,* 55, 120 (1948), and *Soap and Sanitary Chemicals,* 24, 75 (1948).
2. McBain, J. W., and collaborators, in "Colloid Chemistry, Theoretical and Applied," ed., J. Alexander, Vol. V, p. 102, New York, Reinhold Publishing Corp., 1944.
   Hartley, G. S., "Aqueous Solutions of Paraffin-chain Salts," Paris, Hermann et Cie., 1936; *Kolloid-Z.,* 88, 22 (1939).
   Howell and Robinson, *Proc. Roy. Soc. (London),* 155A, 386 (1936).
   ———— and Warne, *Proc. Roy. Soc. (London),* 160A, 440 (1937).
   Lottermoser, A., and Pueschel, *Kolloid-Z.,* 63, 175 (1933).
   ————, and Frotscher, *Kolloid-Beihefte,* 45, 303 (1937).
   Ekwall, *Acta Acad. Aboensis, Math. et Phys.,* (IV), 6, 43, 57, 120, 162 (1927); (VI), 2, 3, (1928); *Z. phys. Chem.,* 161A, 195 (1932); *Kolloid-Z.,* 77, 320 (1936); 80, 77 (1937); 85, 16 (1938); 92, 141 (1940); 94, 42 (1941); 95, 71 (1941).
   Reed, R. N., and Tartar, *J. Am. Chem. Soc.,* 58, 322 (1936).
   Wright, Abbott, Sivertz, and Tartar, H., *J. Am. Chem. Soc.,* 61, 549 (1939).
   Tartar and Wright, *J. Am. Chem. Soc.,* 61, 539 (1939); Tartar and Cadler, *J. Phys. Chem.,* 43, 1173 (1939); Scott, A. B., and Tartar, *J. Am. Chem. Soc.,* 65, 692 (1943); Scott, Tartar, and Lingafelter, *J. Am. Chem. Soc.,* 65, 698 (1943).
   Haffner, Piccione, and Rosenblum, *J. Phys. Chem.,* 46, 622 (1942).
   Ralston, A. W., Hoerr, C. W., and Hoffman, *J. Am. Chem. Soc.,* 63, 2576 (1941); 64, 97, 516, 2824 (1942). Ralston, A. W., and Hoerr, C. W., *J. Am. Chem. Soc.,* 64, 772 (1942).
   Miles, G. D., and Shedlovsky, L., *J. Phys. Chem.,* 48, 57 (1944).
   Dreger, Keim, Miles, G. D., Shedlovsky, L., and Ross, *Ind. Eng. Chem.,* 36, 610 (1944).

Miles, G. D., and Ross, *J. Phys. Chem.*, **48**, 280 (1944).
———, *J. Phys Chem.*, **49**, 71 (1945).
Thiessen and Triebel, *Z. phys. Chem.*, **156A**, 309 (1931).
———, and Spychalski, *Z. phys. Chem.*, **156A**, 435 (1931).
———, and Stauff, J., *Z. phys. Chem.*, **176A**, 397 (1936).
———, *Z. angew. Chem.*, **51**, 318 (1938).
———, Stauff, and Wittstadt, *Z. angew. Chem.*, **49**, 640 (1936).
Stauff, *Naturwissenschaften*, **27**, 213 (1939); *Kolloid-Z.*, **89**, 224 (1939); **96**, 244 (1941).
Hess, K., and Gundermann, *Ber.*, **70B**, 1800 (1937).
———, Philippoff, W., Kiessling, *Kolloid-Z.*, 40 (1939).
Kiessig, H., and Philippoff, *Naturwissenschaften*, **27**, 593 (1939).
———, *Kolloid-Z.*, **96**, 252 (1941); **98**, 213 (1942).
Philippoff, W., *Kolloid-Z.*, **96**, 255 (1941).
3. Brill and Rieder, *Z. angew. Chem.*, **53**, 100 (1940).
Thiessen, *Kolloid-Z.*, **101**, 241 (1942).
4. Chwala, A., and Martina, *Textile-Rundschau*, **2**, 147 (1947).
5. Götte, *Die Chemie*, **57**, 67 (1944).
Aicken, *J. Soc. Dyers Colourists*, **60**, 60, 170 (1944).

EDITOR'S NOTE

The ever-expanding use of synthetic detergents is giving concern to those who control the plants where sewage and trade wastes must be treated. Very stable and well-protected dispersoids formed by use of these detergents may prevent formation of settleable aggregates at any practically obtainable pH. The populations of microorganisms and the activities of their enzymes may be seriously affected. The effect of the detergents in removing or preventing formation of protective films of lime soaps, etc., on sewer pipes, must be considered. See paper by H. H. Goldthorpe, W. H. Hillier, C. Lumb and A. S. C. Lawrence, with discussion, in *Chemistry and Industry*, Oct 1, 1949; also "The Future of Synthetic Detergents in Relation to the Petroleum-Chemical Industry," by A. K. Simcox, *Chemistry and Industry*, March 11, 1950.

THE RELATION BETWEEN CHEMICAL STRUCTURE
AND PERFORMANCE OF DETERGENTS

D. S. P. Roebuck

Nickell Research Labs., Monsanto Chemicals Ltd., Ruabon, N. Wales

SURFACE ACTIVITY (and practically all detergents fall within the category of surface active material) is a property associated with a group of compounds lying between the classifications *soluble* and *insoluble,* whose solutions may pass from true to colloidal state by a small change in composition. This duality of character arises when soluble and insoluble groups are combined in the same molecule. This may be done in two ways: a large insoluble group and a single solubilizing group may be combined, as in the case of the sodium alkyl sulfates, or a series of small units, each containing both soluble and insoluble functions may be combined, as in the polyethylene glycols. As would be expected, the transition from true to colloidal solutions by limited molecular association is only a marked characteristic over a relatively limited range of molecular dimensions. Surface activity, arising from the adsorption of a solute in the surface layer of its solution is yet another manifestation of the duality of character of the capillary-active group of compounds.

Detergency

Detergency is a process of change from a system of interfaces consisting of a surface-contaminant and a contaminant-air interface, to one involving only a surface-air interface, the contaminant having been removed in the course of transition. The transition, brought about by the agency of the detergent solution, must go through several intermediate steps. It will be necessary to establish an interface between the solution and the contaminant, a contaminant-air interface being replaced in the process. The adhesive forces acting between the contaminant and soiled surface must be broken down, the contaminant being displaced into the body of the solution, and an interface of surface-detergent established. Subsequently the detergent-surface interface must be changed by a rinsing

529

process to a surface-solvent interface, which by evaporation will ultimately give rise to a surface-air interface.

The first stage, the replacement of a contaminant-air interface by a contaminant-solution interface, implies that the detergent solution shall wet the contaminated surface. This condition is considered in the well-known theorem, Niemann's triangle, which considers the stability of a liquid drop placed on a surface. If γSA, γLA, γSL are the interfacial tensions associated with solid and air, liquid and air, and solid and liquid interfaces respectively, the condition that the drop shall be in equilibrium on the surface is given by the equation

$$\gamma SA = \gamma LA \cos \theta + \gamma SL$$

where θ is the contact angle. That is, the energy required to replace unit surface-air interface by liquid-air interface $(\gamma SL - \gamma SA)$ is $-\gamma LA \cos \theta$. This has a negative value for θ between 0 and 90°, and a positive value when θ is greater than 90°; hence wetting occurs for values of θ less than 90°. It may be noted that for most effective detergents $\theta = 0$, and their relative wetting efficiencies will depend on the energy gain $(\gamma SL - \gamma SA)$. The critical factor, γSL, is a property determined equally by the nature of the detergent and contaminant.

The breakdown of the surface-contaminant forces, which must follow wetting out of the soiled surface, can involve a range of bond strengths from weak van der Waal's attraction to relatively powerful polar bonds, dependent on the nature of the partners giving rise to the interface. It is to be expected, therefore, that this stage in the detersive action will involve an activation energy which may under some circumstances be considerable. If detersive action is to occur, the reaction:

Surface-contaminant + Detergent → Surface-detergent + Contaminant
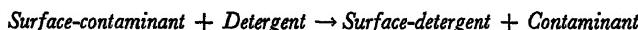
must be energetically favored. At the same time, the surface-detergent complex must not be sufficiently stable to prevent the ultimate removal of the detergent by a rinsing process. Furthermore, it is necessary that the soil removed from the system shall be held in solution or suspension sufficiently tenaciously to prevent its ultimate redeposition on a freshly cleansed surface.

In viewing detersive action in relation to chemical structure, it will be necessary to take all the factors noted above into consideration together with other properties, such as the sensitivity of the aqueous solutions of the detergents to polyvalent ions, which may impose a serious limitation on the utility of an otherwise satisfactory structure.

Wetting and Penetration

In order that effective detersive action may occur, rapid wetting of the contaminated surface is an essential primary condition. It is pertinent, therefore, to consider the phenomena involved in reduction of surface and interfacial tensions of a solvent by a solute, since low interfacial tensions are a necessary condition of ease of wetting.

The early researches of Gibbs revealed that reduction of surface tension of a liquid by a solute is associated with the adsorption of that solute in the liquid surface. This property is exhibited by a large class of compounds having in their structures both hydrophobic and hydrophilic groups. Of this large group, those compounds which are found to have useful detergent properties constitute a smaller group of compounds capable of giving rise to solutions exhibiting colloidal properties. It is necessary, therefore, to consider the structure of the solute in respect to two interrelated phenomena: namely, reduction of interfacial tension and molecular associations giving rise to aggregates of colloidal dimensions. The reduction of interfacial tension is essentially a function of the interfacial layer of material containing an excess of solute over that contained in the interior of the solution, owing to adsorption of the solute in this layer. The mechanism of the adsorption process and the nature of the adsorbed layer are therefore of importance. A consideration of the dynamic aspects of the system serves to indicate the means by which adsorption occurs.

In the interior of the liquid the solvent and solute molecules move chaotically with equal velocities, according to their respective thermal translational energies, in all directions; and uniform concentration of solute exists throughout the system. At the interface, however, the two molecular species are subject to attractive forces directed inwards towards the liquid. These forces are unequal, the solvent molecule suffering a greater pull than the solute molecule with its large hydrophobic group. Initially the ratio of the number of solute to solvent molecules reaching the surface will be determined by the concentration of solute present. The ratio of the numbers of solute to solvent molecules returning from the surface will, however, be lower, as a result of the different attractive forces on the two species. Equilibrium will be reached, therefore, only after a surface excess of solute has been established, so that the numbers of solute molecules reaching and leaving the surface in a given time are equal. An adsorbed film of solute molecules is thus set up. The investigations of Langmuir, Harkins, and others into the structure of adsorbed monolayers has revealed that these films derived from soluble materials are gaseous in nature, unlike the condensed liquid or solid monolayers obtained by spreading insoluble substances on a liquid surface.

Measurement of the colligative property, surface tension,[16] refractive index,[7, 9-11] electrical conductivity,[21, 28] etc., of a surface-active solute, for a range of solutions of various concentrations, reveals a discontinuity in the relationship between the measured property and concentration. This discontinuity is associated with the appearance of colloidal aggregates of solute molecules in the system, and is generally referred to as the critical micelle concentration. It is significant that for concentrations lower than the critical micelle concentration, each successive increment of surface-active agent produces a marked lowering of surface tension, while further additions after the critical micelle concentration has been reached cause very little further diminution of surface tension.

It was initially suggested by Murray [17] and subsequently supported by the studies of Powney and Addison,[19] working with highly purified sodium alkyl sulfates, that reduction of the surface tension of a solvent by a solute capable of giving rise to colloidal solutions was a property arising from the presence of free-solute molecules, and to which the colloidal aggregates contributed to only a minor extent or not at all. More recently, and this work will be referred to again below, Preston [20] has provided an extension of this theory, having obtained experimental evidence that detersive efficiency, as measured by launderometer technique, also depends on the unaggregated solute molecules present in the solution.

We have indicated above that the adsorbed surface excess of solute behaves essentially as a gaseous film, and it is convenient to carry this analogy a little farther. In the case of the two-dimensional gaseous system, we may assume that the solute concentration is analogous to pressure in a true gaseous system in three dimensions, and we may therefore consider the collision frequency between solute molecules in the adsorbed films to be determined by the concentration of molecules. As in a three-dimensional gaseous system, deviations from an ideal gas law will arise. since the collisions between adsorbed solute molecules moving laterally in the adsorbed layer will not be perfectly elastic, and these deviations will be increasingly significant as the concentration increases. Ultimately, a concentration will be reached at which the contact time between colliding molecules will be similar to the time elapsing between one collision and the next, thus initiating the formation of a condensed phase. It seems reasonable to suppose that the onset of molecular association will first appear in the adsorbed surface layer, since this is the region of higher solute concentration. Furthermore, since the association would be expected to arise from interaction of the hydrophobic hydrocarbon portions of the molecule rather than from the similarly charged ionic hydrophilic groups, the hydrophilic property of the associated complex will bear a simple additive relationship to the number of molecules in the complex,

whereas the hydrophobic function will be less than additive, to the extent to which molecular hydrophobe-solvent interface has been replaced by a molecular hydrocarbon-hydrocarbon interface in the process of association. The complex associated structure therefore becomes less hydrophobic and more hydrophilic in nature than its component molecules, and will therefore be withdrawn from the surface layer into the body of the solution. In this way, by making the very reasonable assumption that the onset of micelle formation occurs in the region of maximum solute concentration (the surface), we may readily understand the coincidence of the critical micelle concentration with the attainment of maximum reduction of surface tension, and the unique importance of single-solute molecules in reduction of surface and interfacial tensions.

The mechanism outlined above may be summarized as a cyclic process:

$$
\begin{array}{ccc}
& A & \\
\textit{Free solute molecules} & \rightleftharpoons & \textit{Solute molecules ad-} \\
\textit{in the liquid} & & \textit{sorbed in the surface} \\
& & \\
D \Big\updownarrow & & \Big\updownarrow B \\
& C & \\
\textit{Micelles in the} & \rightleftharpoons & \textit{Associated molecules} \\
\textit{liquid} & & \textit{adsorbed in the sur-} \\
& & \textit{face}
\end{array}
$$
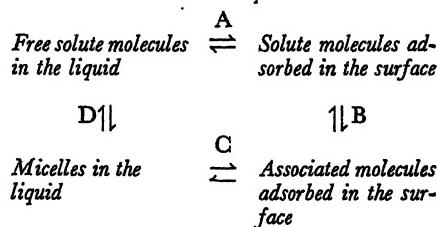
It is apparent that as the hydrophobic portion of the molecule is increased in size, the tendency to adsorption in the surface layer for a given homologous series will also increase, with consequent improvement in efficiency of wetting and penetration; this gain, however, may be offset if simultaneously the tendency to associations also increases. That the opposing factors, surface adsorption and association, are in fact both favored by emlargement of the hydrophobic portion of the molecule is indicated by several studies[3, 5, 7, 24, 26, 28] of the variation of critical micelle concentration * on ascending a homologous series. It is generally found that, for a homologous series, the critical micelle concentration is related to the number of carbon atoms (N) in the hydrocarbon chain by an expression of the form:

$$\log \text{CMC} = k_3 N + k_4$$

where k_3 and k_4 are constants, k_3 being negative, and the expression is limited to a fairly narrow range of carbon atoms.

Having considered the adsorption process at the surface of the solute, it is now necessary to refer briefly to the nature of the associative forces which are responsible for micelle formation. In the case of detergents having a polar hydrophilic group, it is immediately apparent that association must be a function of the hydrophobic portion of the molecule which is remote from the polar influence of the hydrophilic group.

* This is abbreviated to CMC.

Collisions between solute molecules will be effective as associations if the van der Waal's attractive forces at the point of impact of the molecules exceed the polar repulsive forces. This condition can only be satisfied if the molecular weight of the hydrocarbon portion of the molecule is above a definite minimum value which is determined by the nature of the polar head; however, this minimum will rise with progressive branching of the hydrocarbon chain. The influence of the shape of the hydrophobic portion of the molecule has been studied by McBain and Brady[18] and Hartley.[6] These authors have demonstrated that the onset of micellar aggregation is controlled by the ease of packing of the molecules. Thus, a detergent having a straight carbon chain shows a lower CMC than one having a branched chain of the same molecular weight, and the latter, in turn, has a lower CMC than a molecule whose hydrophobic unit is made up of the less readily packed polycyclic hydrocarbon.

It is evident that, from the point of view of wetting and penetration, there will be an optimum balance of hydrophobic and hydrophilic property in the molecule. For a given water solubilizing group the optimum condition will be attained with the length of hydrocarbon chain which gives a maximum extent of adsorption of solute in the surface of the solution, prior to the onset of micellar aggregation. The common hydrophilic groups occurring in the main classes of detergents and wetting agents are the sulfate, carboxylate and sulfonate groups. In the case of the sulfates, the C_8 to C_{18} range of alcohols provide useful starting materials, and the optimum balance of hydrophobic and hydrophilic property is found in the range C_{12} to C_{18}.[18, 25] Mixed lauryl and myricyl alcohols, oleyl alcohol, and stearyl alcohol may all be sulfated to give products of commercial importance as wetting agents. Similarly, the sodium salts of C_{10} to C_{17} aliphatic carboxylic acids provide an important group of detergents embracing the traditional soaps. The alkali stearates (C_{18}) fall outside the group, their value being limited by their low solubility. The stearates are, however, effective wetting agents at elevated temperatures, and the rather more soluble oleates having the same number of carbon atoms are satisfactory detergents at 30°C. A useful range of products, from the point of view of balance of hydrophobic and hydrophilic property, occurs when C_{10} to C_{18} hydrocarbon chains are united with the sulfonic acid grouping.[1, 22] In the case of the so-called nonionic detergents, of which the most important group are those based on the condensation products of ethylene oxide, the hydrophilic unit

$$-O[CH_2CH_2O]_n-CH_2CH_2OH$$
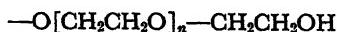
where $n = 5$, may be considered to be approximately equal in solubility power to that of the sulfate group. Some commercial products of this class are listed at the top of facing page:

"Triton" X 100	Diisobutyl phenol condensed with ethylene oxide	$n = 8$ or 9
Polyethylene glycol monolaurate 400	Glycol monolaurate condensed with ethylene oxide	$n = 9$
"Emulphor" 0	Oleylalcohol condensed with ethylene oxide	—

Of the cationic salts, the well-known cetyl piperidinium halides serve as an example.

Before proceeding to review in detail the influence of size and shape of the hydrophobic portions of surface-active molecules, it seems appropriate to consider at this stage the sensitivity of the various types of anionic detergents to precipitation by polyvalent ions. Because of their sensitivity to calcium and magnesium ions, the utility of the traditional carboxylic acid salts was limited in hard-water areas. This stimulated the search for alternative materials, long before present-day economic conditions and foodstuff shortage rendered the use of synthetic detergents in partial replacement of soap imperative. The well-known Turkey Red oil is an example of one of the products of these earlier endeavours.

Since the solute in a detergent solution occurs in two species, free molecules and aggregates, there will evidently be two operative mechanisms by which precipitation by heavy metal ions can occur. If the solubility product of the calcium or magnesium salts of the surface-active acid is low, precipitation of the salts will occur correspondingly, readily involving the unaggregated detergent molecules. The second mechanism pertains to the soap micelles. These colloidal particles owe their stability to an electrical double layer, made up of the negative charges on the ionized polar heads of the molecules and a surrounding cloud of positive ions. The equilibrium of the system depends on the equality of the thermal energy of the surrounding ions, the electrostatic attraction between them and the charged micelle surface. If the charge on the micelle surface is destroyed by the adsorption of polyvalent ions, the stabilization mechanism breaks down, resulting in coalescence of the micelles, and precipitation results.

The sulfates and sulfonates, being salts of strong acids, are not markedly sensitive to calcium or magnesium ions, and may generally be used in hard water. The carboxylates, however, are sensitive, and cannot be used conveniently in hard-water regions. A further disadvantage of the rather weak acid properties of the carboxylic group is a tendency for precipitation to occur if the pH of the solution falls below 10, probably arising both from the low solubility of the single undissociated acid molecules, and from the reduction of the micellar charge as a result of the incomplete dissociation of the polar heads. A variety of chemical reactions have been employed to modify the carboxylic acids in order to eliminate these undesirable properties. "Igepon" T,[2] a detergent par-

ticularly useful in shampoo formulations, is a product of such modification. It is derived by converting oleic acid to its chloride, and reacting the latter with the sodium salt of taurine.

$$C_8H_{17} \cdot CH{=\!=\!=}CH{-}C_7H_{15} \cdot COCl + H_2N \cdot CH_2 \cdot CH_2 \cdot SO_3Na$$
$$\longrightarrow C_8H_{17} \cdot CH{=\!=\!=}CH \cdot C_7H_{15} \cdot CO \cdot NH \cdot CH_2 \cdot CH_2 \cdot SO_3Na$$
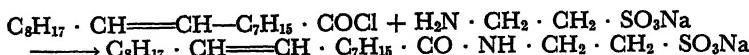
Another modifying process, applied, for example, to stearic acid,[25] involves converting the acid to amide, reacting the amide with ethylene oxide, and esterifying the resulting alcohol with sulfuric acid:

$$C_{17}H_{35} \cdot COOH \rightarrow C_{17}H_{35} \cdot CONH_2 \rightarrow C_{17}H_{35} \cdot CO \cdot NH \cdot CH_2 \cdot CH_2 \cdot OH$$
$$\rightarrow C_{17}H_{35} \cdot CONH \cdot CH_2 \cdot CH_2 \cdot CH_2 \cdot O \cdot SO_3Na$$
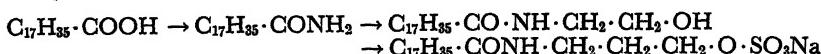
In the previous paragraphs dealing with the order of molecular size of a hydrocarbon group which may be associated with a given hydrophilic group to form a usefully balanced structure, fairly wide ranges of hydrophobe size were indicated. The optimum balance, however, is attained over a considerably narrower range of carbon chain lengths.
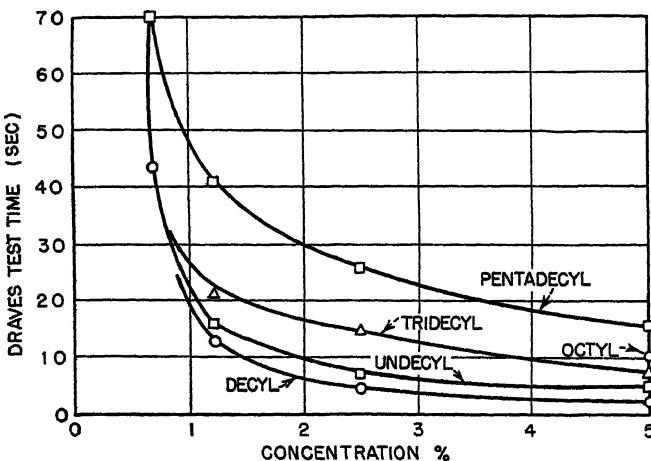
Figure 1. Effect of alkyl chain length and concentrations.

In Figure 1, wetting times as determined by the Draves Clarkson technique, are plotted against concentration for a series of compounds of the type

$$R{-}\langle\!\!\!\bigcirc\!\!\!\rangle{-}SO_3Na$$
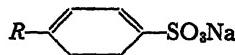
where R varies from C_8 to C_{15} normal paraffin chains. It is apparent from this diagram that optimum chain length is achieved with the C_{10} and C_{11} hydrocarbons. In the case of sodium octyl benzene sulfonate, a reasonable wetting time of 10 seconds is only achieved when the solution contains

5 per cent of wetting agent; 3 per cent of the tridecyl derivative, and between 1 and 2 per cent of the decyl and undecyl derivatives are required to give corresponding wetting times.

Another series of compounds which have been investigated by Caryl,[1] also containing the sulfonate hydrophile, are derived from maleic anhydride by the reaction sequence:

$$\begin{matrix} CH \cdot CO \\ \| \hspace{0.7cm} {\Large >} O + 2R \cdot OH \longrightarrow \\ CH \cdot CO \end{matrix} \qquad \begin{matrix} CH \cdot COOR \\ \| \\ CH \cdot COOR \end{matrix}$$

$$\begin{matrix} CH \cdot COOR \\ \| \hspace{1cm} + NaHSO_3 \longrightarrow \\ CH \cdot COOR \end{matrix} \qquad \begin{matrix} CH_2 \cdot COOR \\ | \\ CH \cdot COOR \\ | \\ SO_3Na \end{matrix}$$
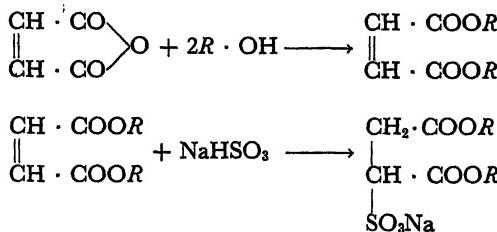
The reaction shows a similar narrowness in the range of size of the hydrocarbon radical R, giving optimum wetting power. Wetting power was found to increase over the range $R = C_3$ to C_7, and decrease over the range C_7 to C_{10}.

In the two examples which have just been considered, the variation in hydrophobic function of the molecule has arisen through a change in the length of a hydrocarbon chain. It is interesting, therefore, to consider an example where a chemical rather than physical characteristic (dimension) is operative. In the case of the sodium butyl biphenyl sulfonates,
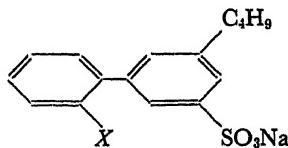
the following Draves test times were obtained:

Concentration (%)	X = H	Time (sec) X = OH	X = Cl
0.5	11.6	10.9	4
0.25	23.1	38.7	7
0.125	78.4	71.8	21
0.063	171.8	—	113

These figures illustrate that reduction of the hydrophobic properties of the molecule by including a second hydrophile in the benzene ring system reduces the wetting efficiency of the product. Inclusion of the halogen substituent produces a marked improvement in wetting power, possibly because the marked polar character of the chlorine to carbon bond tends to inhibit micellar aggregation.

Destruction of the Adhesive Forces Between Soil and Contaminated Surface

The adhesive forces acting between the soil and the surface may be polar or nonpolar in character, and in the case of hydrocarbon soils, polar bonds may be progressively generated by oxidation of the hydrocarbon contaminant. With nonpolar soils, the hydrocarbon portion of the detergent molecule adsorbed at the interface would be expected to become adsorbed on the contaminant surface, thus producing a complex in which the hydrophobic function associated with the hydrophilic group is considerably greater than in the original detergent molecule. Such a complex would be expected to have a much lower CMC than the original detergent, and it may therefore tend to be withdrawn into the body of the solvent as a colloidal aggregate, possibly after further association stops.

Where the soil is attached to the surface by polar forces, ion-exchange phenomena may play an important part in detersive action. This aspect of the problem has recently been reviewed by McBain.[12] He demonstrated that sulfate, silicate, and phosphate ions are all effective in removing dyes from cotton and glass. These ions are often found in the inorganic builders employed in the blending of commercial detergents.

There are two types of reactions in which these ions might be expected to participate. Where a detergent molecule has formed an associative complex with the hydrocarbon portion of a polar soil molecule, attached by its polar head to a surface, a displacement type of reaction in which surface contaminant bond was replaced by a surface-builder bond would greatly facilitate the passage of the soil molecule into the liquid. Alternatively, where a detergent molecule has displaced the soil molecule from the surface, it in turn might be liberated from the surface by a builder ion, thus maintaining effective detergent concentration at the interface.

The mode of operation of inorganic anions may also depend on the occurrence of phenomena related to the appearance of minima in the surface tension and concentration curves of typical detergent ions, thought to arise as a result of the influence of the anions on the dimensions of the stabilizing electrical double layer associated with the adsorbed surface layer of capillary-active solute.[16, 22] Harris[4] has pointed out that in the case of sodium dodecyl benzene sulfonate, launderometer measurements show that the efficiency of soil removal parallels the lowering of surface tension occurring on the addition of neutral or alkaline electrolytes.

A type of mechanism which may in some respects be considered as a model for the experiment described above has been reported in a recent paper by Hutchinson.[8] This paper discusses the formation of mixed

adsorption films of an alcohol of low solubility and typical detergent anions. It was shown that octyl alcohol was the main constituent of the adsorption layer in the presence of sodium dodecyl sulfate, exerting a higher surface pressure than the detergent molecule when the sulfate was present in concentrations lower than the CMC. As the sulfate concentration was raised to the CMC, the effect of the octyl alcohol was reduced, possibly indicating that the alcohol becomes solubilized and is carried to the interior of the solution in micelles.

It has been suggested that single adsorbed molecules in the surface layer of a solution of detergent are responsible for the reduction of surface tension of the solute, and that the effectiveness of the detergent in this respect is limited by the tendency of the molecules to associate, a property determined by the size and shape of the hydrophobic portion of the molecule. Similarly, association of the adsorbed solute molecules with a contaminant has been considered as a primary step in the breakdown of surface-soil cohesive forces. This appears to be borne out by the work of Preston.

Launderometer measurements with solutions of varying concentrations of sodium lauryl sulfate showed that detersive efficiency closely paralleled reduction in surface tension, only small increases in efficiency being obtained when the detergent concentration was increased beyond the critical micelle concentration; a critical washing concentration was observed coincident with the critical micelle concentration. Furthermore, Preston's data for molar critical washing concentrations indicated that this was a function of the number of carbon atoms in the hydrocarbon chain and was independent of the nature of the hydrophilic group.

Solubilization of the Liberated Soil

The means by which the liberated soil is prevented from being redeposited on the clean surface must now be considered. It has been suggested that the particles of soil are removed from the surface as association complexes with detergent molecules and are incorporated in micelles. There is little doubt that micelles are involved in the solubilization process. It has been demonstrated, for example,[14, 15] that hydrocarbons such as propylene, methyl cyclopentane, isobutane, etc., are solubilized in water by the addition of soaps, the process depending on the formation of colloidal aggregates. There are not yet sufficient data to indicate the precise structure of the colloidal aggregates, but this will certainly affect the solubilizing efficiency of the micelle. It seems apparent that two factors are important, namely, the free space within a micelle, which is available to the contaminant, and the expansion of the surface, with

consequent reduction of the charge density on the micelle, which can be tolerated before the system becomes unstable and coalescence occurs.

A recent paper by Vetter,[27] on the properties of solutions of the sodium salt of sulfonated di-2-hexyl succinate ("Aerosol" MA), seems to support Hartley's early suggestion of a spherical micelle, with the hydrocarbon chains of the colloidal electrolyte directed towards the center, except that the micellar volume appears to be greater than that postulated by Hartley. This indicates a greater penetration of the aqueous phase towards the paraffinoid center of the micelle. At the same time there is a considerable body of x-ray data in support of the existence of lamellar micelles. Much work remains to be done to elucidate the factors which determine the type of micelle formed, and how much removed soil the latter will be capable of accommodating.

REFERENCES

1. Caryl, C. R., *Ind. Eng. Chem.*, **33**, 731 (1941).
2. Guenther, F., U.S. Patent 1,932, 180 (1933).
3. Harkins, W. D., *J. Am. Chem. Soc.*, **69**, 14, 28 (1947).
4. Harris, J. C., *Oil & Soap*, **33** (4), 101 (1946).
5. Hartley, G. S., *Kolloid-Z.*, **88**, 22 (1939).
6. ———, *Trans. Faraday Soc.*, **37**, 130 (1941).
7. Hess, K., Philippoff, W., and Kiessig, H., *Kolloid-Z.*, **88**, 40 (1939).
8. Hutchinson, E., *J. Coll. Sci.*, **3**, 413 (1948).
9. Kleevens, H. B., *J. Chem. Phys.*, **14**, 567 (1946).
10. ———, *J. Coll. Sci.*, **2**, 301 (1947).
11. ———, *J. Phys. Coll. Chem.*, **52**, 130 (1948).
12. McBain, J. W., in "Advances in Colloid Science," eds. E. O. Kraemer, *et al.*, Vol. I, p. 99, New York, Interscience Publishers, Inc., 1942.
13. ———, and Brady, A. P., *J. Am. Chem. Soc.*, **65**, 2072 (1943).
14. ———, and O'Connor, J. J., *J. Am. Chem. Soc.*, **62**, 2855 (1940).
15. ———, and ———, *J. Am. Chem. Soc.*, **63**, 875 (1941).
16. Miles, G. D., and Shedlovsky, L., *J. Phys. Chem.*, **48**, 57 (1944).
17. Murray, R. C., *Trans. Faraday Soc.*, **31**, 206 (1935).
18. Palmer, R. C., *J. Soc. Chem. Ind.*, **60**, 56 (1941).
19. Powney, J., and Addison, C. C., *Trans. Faraday Soc.*, **33**, 1243 (1937).
20. Preston, W. C., *J. Phys. Coll. Chem.*, **52**, 84 (1948).
21. Ralston, A. W., and Hoerr, C. W., *J. Am. Chem. Soc.*, **64**, 772 (1942).
22. Reed, R. N., and Tartar, H. V., *J. Am. Chem. Soc.*, **58**, 322 (1936).
23. Reichenberg, D., *Trans. Faraday Soc.*, **44**, 467 (1947).
24. Scott, A. B., and Tartar, H. V., *J. Am. Chem. Soc.*, **65**, 692 (1943).
25. Snell, F. D., *Ind. Eng. Chem.*, **35**, 107 (1943).
26. Stauff, J., *Z. physik. Chem.*, **183A**, 55 (1938).
27. Vetter, R. J., *J. Phys. Coll. Chem.*, **51**, 262 (1947).
28. Wright, K. A., Abbot, A. D., Sivertz, V., and Tartar, H. V., *J. Am. Chem. Soc.*, **61**, 549 (1939).

GRAHAM'S SALT AND ITS USES

Thomas Hill Daugherty

Director of Research, Calgon, Inc., Pittsburgh, Pa.

GRAHAM'S SALT is sodium metaphosphate glass. For nearly a century after its discovery, it remained a laboratory curiosity with only minor uses in analytical chemistry. Then a series of discoveries was made which lifted it into the realm of important basic chemicals. Today it is vital in many large industrial operations and has found numerous applications in the home. The first commercial production in 1929-30 consisted of 2000 pounds. The next year 1,100,000 pounds was produced. In 1947, the total production in the United States was approximately 60,000,000 pounds, with additional quantities produced in other countries. This summary will be limited to a presentation of the basic chemistry of Graham's salt and its homologs, a brief review of the history pertaining to the discovery of their several properties and a descriptive listing of the important applications which are being made of these properties.

Historical

Berzelius[1] in 1816 prepared monosodium orthophosphate which he dried at red heat before analyzing for Na_2O and P_2O_5 without apparently recognizing the formation of a new chemical species in so doing. Proust,[2] in 1820, prepared a glass by heating microcosmic salt, $NaNH_4HPO_4$, but thought it to be a sodium orthophosphate resulting from the loss of ammonia. Thomas Graham, father of colloid chemistry, found[3] in 1833 that by heating "biphosphate of soda," NaH_2PO_4, in stages he could produce four new and different substances. Three of these were crystalline, but the fourth was a glass which he named "metaphosphate of soda." This glass has become known in the literature as "Graham's salt" and as "sodium hexametaphosphate."

Graham's salt has the molecular formula $(NaPO_3)_x$ or an empirical molecular ratio of $1Na_2O:1P_2O_5$. It is conveniently prepared by roasting monosodium dihydrogen orthophosphate (NaH_2PO_4) or disodium dihy-

drogen pyrophosphate ($Na_2H_2P_2O_7$) to drive off the water of hydration and most of the water of constitution, fusing the residue (m.p. 627.6° C) [4] and rapidly chilling the melt. In the familiar metaphosphate bead test developed by Emerson,[5] glassy sodium metaphosphate is formed from microcosmic salt ($NaNH_4HPO_4$) in a similar manner. The clear and colorless glass which results is readily soluble in water to form an unbuffered, nearly neutral solution. The glass is best dissolved by being suspended just under the surface of the water or else by slow addition of the powdered form to the water with stirring. The solubility in liquids other than water is negligible. As concentration of the solution in water is increased, the viscosity increases. At 82°F, a solution containing 50 per cent by weight has a viscosity of approximately 102 centipoises and one containing 70 per cent by weight has a viscosity of approximately 4220 centipoises. As the temperature is increased, the viscosity decreases. At 110°F, the two solutions cited have viscosities of approximately 60 and 1740 centipoises respectively.

This glass was designated as a "hexametaphosphate" by Fleitmann,[6] solely from analytical evidence. Other terms implying degree of polymeric association, such as "tetraphosphate," [7] "septaphosphate" [8] and "decaphosphate," [7] have also been applied to sodium phosphate glasses of various $Na_2O:P_2O_5$ ratios. However, molecular weight determinations from end-group titrations,[9] dialysis,[10,11] conductivity [12] and ultracentrifuge sedimentation,[13,14] reveal values from 1000 to 17200, depending upon the method employed, the $Na_2O:P_2O_5$ ratio, the previous thermal history of the glass and the residual water content. These measurements do not substantiate the existence of 4, 6, 7, 10 or other integral number of units in the polymeric molecules of these glasses but instead indicate a variety of chain lengths, the average of which for a given sample depends upon the thermal treatment in its preparation and to some extent upon its $Na_2O:P_2O_5$ ratio. Reviews of the molecularly dehydrated or condensed sodium phosphates have been made by Terrey,[15] Karbe and Jander,[10] Quimby,[16] and "Thorpe's Dictionary of Applied Chemistry." [17] In particular, Partridge [18] has discussed the tangled terminology of the complex phosphates and recommends that a sodium phosphate glass be designated by its $Na_2O:P_2O_5$ ratio. In his system, Graham's salt would be simply "sodium (1:1) phosphate glass."

The earliest property of a metaphosphate to be discovered which eventually resulted in practical applications of Graham's salt was the precipitation of albumin. Engelhart [19] in 1825, before the metaphosphates were recognized as such, found that phosphoric acid precipitated the albumin of blood serum even after a thousandfold dilution of the serum. Berzelius [20] in 1827, however, could not obtain an albumin precipitate with his phosphoric acid. Then he reported [21] in 1828, after

working with Engelhart, that Engelhart strongly evaporated the phosphoric acid (thus probably producing metaphosphoric acid), after which it would precipitate albumin when first dissolved but lost the property after several days' standing in solution. Graham[3] in 1833 employed albumin as a reagent to distinguish the acid of his "metaphosphate of soda," which precipitated albumin, from pyrophosphoric and orthophosphoric acids, which did not precipitate albumin. Aside from applications in analytical chemistry, no use was made of this property until comparatively recently.

The solubilizing action of molten metaphosphates on metallic oxides was first clearly reported by Berzelius[22] in 1844 in his work on applications of the blowpipe, although it was known earlier that monosodium phosphates fused to transparent beads. Emerson[5] in 1866 and Rose[23] in 1867 extended the knowledge of analysis by means of metaphosphate beads and Wallroth[24] in 1883 quantitatively examined the metal salts which crystallized during cooling. Except in the fields of analytical chemistry and mineralogy, this application of the glassy phosphates also has not been employed industrially until comparatively recently.

The deflocculating action of Graham's salt on suspended solids was first reported by Feldenheimer[25] in 1922, who deflocculated clay suspensions with pyrogenic derivatives of orthophosphoric acid and their salts, including the metaphosphates. Chwala[26] in 1929 extended this application by dispersing water-insoluble arsenates and phosphates with "water poor" (i.e., molecularly dehydrated) arsenates and phosphates, including the metaphosphates. The separation and beneficiation of minerals by the assistance of Graham's salt is a related application. It was first reported by Ellis[27] in 1922 who, in concentrating metalliferous ores by flotation with electrolytes producing multivalent ions, mentions the "hexavalent" anions from "hexametaphosphate" and "tetraphosphate." Later, Rose and MacDonald[28] employed metaphosphates, as well as metaphosphites, meta-arsenates and meta-arsenites, for the flotation separation of minerals, and Keck, Eggleston and Lowry[29] employed sodium metaphosphate as activator for the flotation of hematite.

One of the most important characteristics of Graham's salt from the standpoint of commercial application is sequestration or the prevention of precipitation of multivalent cations by complex ion formation. The term is defined by Daugherty[30] as "the reduction of the concentration of a multivalent positive ion in solution by combination with a negative ion to form a complex negative ion, to the extent that the remaining concentration of the multivalent positive ion is insufficient to be precipitated by a given negative ion with which it has a low solubility product constant. The sequestration value is a stoichiometri-

cal ratio or weight relationship expressed as the quantity of a sequestering agent required to capture a unit quantity of a multivalent positive ion and form with it a complex negative ion which is stable against precipitation by a given precipitant for the multivalent positive ion." The calcium value of Andress and Wüst [31] is the sequestration value in the case where the multivalent cation is calcium and the precipitant is the negative ion of the sequestering agent itself; thus, the calcium value is the weight ratio of a given sequestering agent to calcium required to dissolve the initial calcium precipitate of the sequestering agent.

In 1929, Hall and Jackson [32] desired to introduce phosphate into boiler water by means of the feedwater but found that orthophosphates precipitated the calcium in the feedwater, thereby plugging the feedline. They then considered the condensed phosphates but were unable to obtain a single pound of Graham's salt on inquiry from many sources and were forced to manufacture the requirement for their test. The experiment was successful but not until 1932 did Hall [33] make the greater discovery that Graham's salt softened the water against precipitation of the hardness constituents by soap. A similar action was obtained with other bi- and trivalent ions and with other precipitants than soap.

Hall and Jackson's application of Graham's salt to boiler water conditioning resulted in the first commercial production of this chemical and Hall's discovery of sequestration by Graham's salt greatly expanded the commercial production. It is historically significant that Rose [34] in 1849 observed the initial precipitation of alkaline earth metaphosphates by Graham's salt and the re-solution of these precipitates upon addition of excess Graham's salt. Also Scheerer [35] in 1858 reported that his assistant, Rube, had found Graham's salt to interfere with the precipitation of barium as the sulfate and of barium, strontium, and calcium as the carbonate. Practically, these findings lay dormant in the realm of analytical chemistry for some eighty years until Hall uncovered the phenomenon of sequestration.

The action of very low concentrations of Graham's salt in preventing the precipitation of $CaCO_3$ from hard waters is another characteristic which has found wide commercial application. This action has become commonly known as "threshold treatment," which name has a dual significance since it applies to arresting the precipitation of $CaCO_3$ at the threshold of crystallization and also signifies concentrations at the threshold of measurement and control. The threshold effect of sodium metaphosphate glass was first reported by Rosenstein [36] who in 1935 overcame the precipitation of $CaCO_3$ in irrigating water, which occurred when NH_3 was added as a fertilizer, by first introducing a few

parts per million of Graham's salt. The effect is discussed in detail by Hatch and Rice,[37] Rice and Partridge,[38] Reitemeier and Buehrer,[39] and Fink and Richardson.[40]

Threshold treatment of water with Graham's salt has also been found to protect the steel of pipe lines and equipment from corrosion by the formation of a protective film. This action is widely employed for corrosion protection alone and in conjunction with scale prevention. It was first reported by Hatch and Rice [41] in 1940 and is dealt with in the report by Rice and Hatch.[42] About the same time, Rice [43] found that Graham's salt stabilized dissolved iron in water and this action is widely employed to overcome "red water" troubles in municipal supplies.

Commercial Application of Graham's Salt

Commercially, the true Graham's salt, sodium (1:1) phosphate glass, is not produced because it is comparatively slow to dissolve and produces an acid solution. Instead, current practice is to manufacture glassy products of higher $Na_2O:P_2O_5$ ratios, these ranging from 1.1:1 to approximately 1.4:1. These glasses are rapidly soluble in water and produce solutions ranging from neutral to alkaline. They are made in various physical forms, including platelets resembling broken window glass, coarse irregular lumps, beads, coarse powder and fine powder. For the purpose of describing the uses of these modern adaptations of Graham's salt, the first commercial form, Calgon,[44] which is sodium (1.1:1) phosphate glass, will be employed.

Protein Precipitation. The action of acid solutions of the glassy phosphates in precipitating albumin has led to industrial applications involving the reaction with other proteins, such as in the pretanning of hides prior to vegetable tanning. This application is based upon the fundamental work of Wilson,[45] who reported the rapidity of the initial combination of Calgon with hide protein. Subsequently it was found that tannin penetration into the hide was accelerated by the phosphate pretanning and that a stronger and more plump leather resulted. The property is also employed in the deproteinization of products such as sugar, milk, and blood and in the removal of albuminoids from therapeutical fluids such as sera and rattlesnake venom. Interaction with proteins produces emulsification in the processing of cheese and stabilization of wine, lactalbumin and casein glues. Pectin is freed from the proteins naturally occurring with it in apple pomace, citrus wastes and sugar beet residues by metaphosphate treatment. Removal of proteins by metaphosphate precipitation is of increasing importance in disposal or re-use of waste water.

Fluxing Action. Comparatively little practical use has been made of the fluxing or solubilizing action of molten sodium metaphosphate on metallic oxides and other minerals. This property is utilized, however, in removing undesirable impurities from the surface of molten metals, such as aluminum, before pouring. Also, metaphosphate glass serves as the binder in the sintering of mica flakes to form electrical insulators. The purification of carbides in molten metaphosphate and the incorporation of sodium metaphosphate as a constituent of ceramic bodies and glazes have been reported to be satisfactory applications but are not employed commercially so far as is known.

Dispersion. The dispersion of pigments by means of sodium metaphosphate glass is widely employed in the preparation of paper coatings. Relatively insoluble substances, such as clay, lithopone, calcium carbonate, titanium dioxide and barium sulfate, are readily dispersed in water, the slurries produced being characterized by low viscosities at high solids concentrations and by good suspension of the pigment particles. Oil-well drilling muds, composed of bentonitic (swelling) clays, formation (nonswelling) clays and weighting agents such as barium sulfate and iron oxide, are effectively dispersed with low concentrations of sodium metaphosphate to produce the fluidity requisite for recirculation and screening. Likewise, ceramic slips and glaze slurries containing low concentrations of glassy phosphate remain fluid at comparatively high solids concentrations. Pitch control in the manufacture of groundwood paper and increasing the porosity of plugged formations in water wells are other uses of this dispersive property. Reduction of viscosity and consequent improvement in ease of application are obtained in water base paints by dispersion of the pigments with metaphosphates. The deflocculating action of the metaphosphates is advantageously employed in the flotation separation of many types of ores from admixed base minerals. The creaminess and consistency of wax, polish and similar emulsions containing such constituents as grease, casein and latex are considerably improved by addition of glassy phosphates during the mixing operation.

Sequestration. The ability of glassy sodium phosphate to tie up bi- and trivalent metal ions against precipitation by negative ions with which they normally form insoluble compounds is widely employed in operations where precipitation is harmful or undesirable. Such an operation is washing. With soap as the detergent, lime and magnesium soaps precipitate (which wastes soap) to form a scum which deposits on the articles being washed. The sequestration of calcium and magnesium ions by these phosphates completely prevents this precipitation and thus makes possible a better and more efficient washing. In the home, the principal washing operations which can be improved by the use of

non-scum-forming detergent solutions are: bathing, laundering, dish-washing, window washing, floor and wall washing and car washing. Similar benefits are obtainable by the use of a sequestering agent with synthetic detergents less sensitive than soap to precipitation by calcium and magnesium ions, due to solubilization and to dispersion of the constituents of the soil load.

In industry, most of the home uses find employment on a larger scale. Also many other types of precipitants than soap are encountered. Thus the phenomenon of sequestration is advantageously employed in kier boiling cotton, in scouring and bleaching textiles, as a leveler in dye baths, in degumming silk and in degreasing wool. It is used in pulp digestion and for washing paper mill felts. It is also used in dairy cleaning and for the washing of many diverse materials, such as sand, including greensand; bauxite; clays, including zeolitic clays; bottles and other glassware; bricks, walls and terra cotta (but not marble); metals, for removal of soot, grease and oil prior to plating or enameling; fruits and vegetables, to remove insecticidal spray residues; filter cloths; sponges; and sharkskins. The blanching of vegetables, such as peas, in water containing metaphosphate produces a tenderization of the skins, presumably by interaction with the alkaline-earth pectates.

Sequestration with the glassy phosphates is employed to protect emulsions where the soap or other emulsifying agent would be precipitated by the hardness ions in water, typical emulsions being insecticidial sprays, shampoos and cosmetic creams. Also it serves to produce a "soft curd" in milk and milk products. Where presence of multivalent positive ions produces a catalytic effect, a sequestering agent serves as an anti-catalyst by inactivating such ions, examples being stabilization of flavors and odors and prevention of rancidity in food products and the stabilization of hydrogen peroxide solutions. In the water washing of contaminated oils and solvents, the glassy phosphates facilitate the clarification.

Threshold Effect. Low concentrations, of the order of a few parts per million, of sodium phosphate glass are employed in municipal and in home water systems to prevent deposition of scale and to minimize corrosion in piping and in hot water heaters. When added to an iron- or manganese-containing water before oxidation occurs, as when fed to a well water by a drop tube in the well casing or at the pump intake, the glassy phosphate also prevents the appearance of "red water" or "black water" by stabilizing the subsequently oxidized iron or manganese against precipitation. The same effects are obtained with threshold treatment in industrial water systems, principally in once-through and in open recirculatory cooling systems. Scale deposition in evaporating pans is also controlled by threshold treatment. Thus heat exchange surfaces are kept free from the insulating effect of scale and corrosion

products. Where water is passed through granular beds, such as in sand or activated carbon filtration, in zeolite softening, or in returning water to formations (disposal of waste water, repressuring oil fields), the tendency to coat the granules and to plug the interstices by precipitation of calcium and magnesium salts and of hydrous oxides is overcome by the presence of minute concentrations of glassy sodium phosphate. This effect is also widely employed in the prevention of $CaCO_3$ deposition during the deinking of paper.

Miscellaneous Uses. In addition to the properties of Graham's salt which have been cited, several other properties common to the glassy sodium phosphates are known which either have been put to practical use only to a minor extent or are undergoing extensive investigation now. Such properties are: the water binding action, which enables acetic anhydride to be made from acetic acid; the catalytic action, which permits the preparation of organic products from CO (such as acetyl chloride from CO and CH_3Cl), the polymerization of hydrocarbons (such as olefins) and the hydration of olefins to alcohols; and the adsorption action on metal surfaces, which plays a part in the corrosion protection previously mentioned and which is being developed for the protection of metal surfaces under paint or enamel coatings.

Furthermore, a large number of uses for the glassy sodium phosphates based upon the characteristic properties cited are described in the literature but are not specifically mentioned in this summary because they have not been applied thus far beyond the laboratory or pilot-plant stages. For example, very small quantities of this type of phosphate can serve to extend the setting time and improve the mechanical properties of cement and of various plasters. In particular, the multitude of references and claims in the patent literature pertaining to these compounds presages many new uses for the legacy of Thomas Graham.

REFERENCES

1. Berzelius, J., *Ann. Physik*, **54**, 31-55 (1816).
2. Proust, J. L., *Ann. chim. phys.*, 14 (2), 257-88 (1820).
3. Graham, T., *Phil. Trans. Roy Soc. (London)*, 123, 253-84 (1833).
4. Morey, G. W., and Ingerson, E., *Am. J. Sci.*, 242, 1-6 (1944).
5. Emerson, G. H., *Proc. Am. Acad. Arts Sci.*, 6, 476-94 (1866).
6. Fleitmann, T., *Ann. Physik,* 78, 233-60, 338-66 (1849).
7. ———, and Henneberg, W., *Liebig's Ann. Chem.*, 65, 304-34 (1848).
8. Hanlon, R. T., Steffen, A. J., Rohlich, G. A., and Kessler, L. H., *Ind. Eng. Chem.*, 37, 724-35 (1945).
9. Samuelson, O., *Svensk Kem. Tid.*, 56, 343-48 (1944).
10. Karbe, K., and Jander, G., *Koll. Beihefte*, 54, 1-146 (1942).
11. Teichert, W., and Rinman, K., *Acta Chem. Scand.*, 2, 225-51 (1948).
12. Davies, C. W., and Monk, C. B., *J. Chem. Soc.*, 413-29 (1949).
13. Lamm, O., and Malmgren, H., *Z. anorg. Chem.*, 245, 103-120 (1940).
14. ———, *Arkiv. Kemi. Mineral. Geol.*, 17A (25), 1-27 (1944).

15. Terrey, H., *Ann. Reports Chem. Soc. (London)*, 34, 115-22 (1937).
16. Quimby, O. T., *Chem. Revs.*, 40, 141-79 (1947).
17. "Thorpe's Dictionary of Applied Chemistry," Vol. 9, pp. 502-14 (1949).
18. Partridge, E. P., *Chem. Eng. News*, 27, 214-7 (1949).
19. Engelhart, F., *Kastner's Archiv. ges. Natur.*, 6, 337-64 (1825).
20. Berzelius, J., *Jahr. Fort. Phys. Wis.*, 6, 275-9 (1827).
21. ———, *Jahr. Fort. Phys. Wis.*, 7, 117-9 (1828).
22. ———, "Anwendung des Lothrohres in der Chemie und Mineralogie," 4th ed., 1844.
23. Rose, G., *J. Prakt. Chem.*, 101 [1], 217-35; 102 [1], 385-99 (1867).
24. Wallroth, K. A., *Ofversigt af Knogl. Vetenskaps-Akademiens Forhandlinger*, 40, 21-45 (1883).
25. Feldenheimer, W., U. S. Patent 1,438,588 (1922).
26. Chwala, A., U. S. Patent 1,728,622 (1929); Germ. Patent 478,190 (1929); Germ. Patent 504,598 (1930).
27. Ellis, R., U. S. Patent 1,425,185 and 1,425,186 (1922).
28. Rose, E. H., and MacDonald, W. T., U. S. Patent 2,040,187 (1936).
29. Keck, W. E., Eggleston, G. C., and Lowry, W. W., *Am. Inst. Min. Met. Eng.*, Tech. Pub. No. 763, 1-24 (1937).
30. Daugherty, T. H., *J. Chem. Ed.*, 25, 482-8 (1948).
31. Andress, K. R., and Wüst, K., *Z. anorg. allgem. Chem.*, 237, 113-31 (1938).
32. Hall, R. E., and Jackson, H. A., U. S. Patent 1,903,041 (1933).
33. ———, U. S. Patent 1,956,515 (1934); Patent Reissue 19,719 (1935).
34. Rose, H., *Ann. Physik*, 76, 1-13 (1849); *Phil. Mag.*, 34 (3), 321-40 (1849).
35. Scheerer, T., *J. prakt. Chem.*, 75, 113-6 (1858).
36. Rosenstein, L., U. S. Patent 2,038,316 (1936); Patent Reissue 20,360 (1937); Patent Reissue 20,754 (1938).
37. Hatch, G. B., and Rice, O., *Ind. Eng. Chem.*, 31, 51-7 (1939).
38. Rice, O., and Partridge, E. P., *Ind. Eng. Chem.*, 31, 58-63 (1939).
39. Reitemeier, R. F., and Buehrer, T. F., *J. Phys. Chem.*, 44, 35-74 (1940).
40. Fink, G. J., and Richardson, H. H., U. S. Patent 2,358,222 (1944).
41. Hatch, G. B., and Rice, O., *Ind Eng. Chem.*, 32, 1572-9 (1940).
42. Rice, O., and Hatch, G. B., U. S. Patent 2,337,856 (1943).
43. ———, U. S. Patent 2,304,850 (1942).
44. Partridge, E. P., and Texter, C. R., *Chem. Eng. News*, 27, 840-3 (1949).
45. Wilson, J. A., *J. Am. Leather Chem. Assoc.*, 32, 113-31 (1937).

COSMETICS

Paul G. Lauffer

Chief Chemist, George W. Luft Co., Inc., N. Y.

COSMETICS ARE preparations designed to enhance beauty by external application. This result may be obtained by accentuating or supplying desirable features (make-up preparations); by hiding, abolishing or minimizing undesired features (some make-up and treatment items); or by counteracting the effects of age, trauma or disease (either by make-up or treatment). The preparations used divide themselves naturally into two classes; the make-up preparations, which are applied in order to cause temporary improvement in appearance, and the treatment items, which cause no immediate visible change but which are applied repeatedly for the purpose of bringing about a gradual beautifying effect.

Tissues in Cosmetic Application

The skin, hair and nails are the tissues subject to cosmetic care. Since they form the covering of the body, they must be relatively inert and resistant to agents regularly encountered. They function to protect the more delicate tissues from the vicissitudes of the external world. They therefore form an almost ideal base for cosmetics of the make-up type, upon which rouges, powders, etc. can be applied repeatedly without damage. They offer resistance, however, to "treatments" designed to affect them more fundamentally. The discovery of effective and harmless means of invigorating these resistant tissues is the goal of cosmetic chemistry.

Structure of the Skin.[*] The skin is composed of a number of layers, varying progressively in nature from the outermost layer of dry, tough, scaly, dead cells, to the innermost layer with its elastic and collagenous fibers, blood vessels, lymph spaces, nerves, sebaceous and sweat glands, and muscles. The inside layer is an actively living metabolizing tissue

---

[*] For a more detailed description of skin structure, with special emphasis on irritation, see Combes, F. C., in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. 6, pp. 746-81, New York, Reinhold Publishing Corp., 1946.

comparable to deeper tissues; the outside layer is highly differentiated, inert and dead, continuously shed and replaced from below. The composition and properties of the skin therefore differ tremendously according to the layer involved.

The *dermis,* or cutis, forming the two innermost layers of the skin, is mesodermal in origin. Its inner layer contains a lattice of fibers; its outer layer has papillary projections extending into the epidermis. The *epidermis* has at least five distinguishable layers. The deepest (germinal or basal layer) is a single layer of cells, very active and continually proliferating to furnish new cells which push up the older ones in the overlying layers. The second layer (prickle cells) has its cells connected by webs or "prickles" of protoplasm reaching in all directions. These cells are nourished by lymph in the intercellular spaces which connect with lymph ducts in the dermis. The third layer (*stratum granulosum*) becomes more specialized; the cells are flattened and contain horny keratohyalin, a keratin precursor. Next comes the *stratum lucidum,* translucent cells without nuclei, containing eleidin. The outside layer (*stratum corneum*) consists of several layers of hard, horny cells in which the eleidin has changed to keratin. The total thickness of the skin ranges from 0.5 to 4.0 mm.

The outer, horny layer varies in thickness on different parts of the body, and among individuals. The thickness of this layer has some bearing on resistance of the skin to penetration of liquids and solutes, an important consideration for cosmetic practice.

Special structures, important for their part in absorbing oils and oil-soluble materials, are the sebaceous glands, situated in the dermis. The sweat glands, in the subcutaneous tissue, appear more important for excretion than for absorption, but they do seem to afford channels for absorption of aqueous solutions.

Structure of Hair and Nails. The hair and nails are composed largely of keratins, similar to those in the horny layer of the skin. Each hair grows from a hair follicle, formed by involution of the epidermis, which contains the hair bulb, where the cells which are to form the hair shaft proliferate, pushing the shaft outward. The core, or medulla of the hair, is formed of polyhedral cells containing some eleidin. Around the core is the cortex, forming the bulk of the hair and composed of long cells fused into fibers and containing the pigment. On the outside is a single layer of flat cuticle cells, overlapping like shingles. The pigment of the hair is melanin, the same as that of the skin.

The nails grow outward from the nail matrix at the proximal end. They are not distinctly differentiated into layers, but in thick nails three layers may be recognized. The surface is hard and smooth, with longi-

tuginal ridges of varying height, and forms a fairly inert base for mechanical or chemical treatment.

Perspiration. The normal human skin is always covered by the secretion of the sweat glands (more or less evaporated, according to atmospheric conditions) and that of the sebaceous glands. Sweat contains about 0.7 per cent sodium chloride, and 0.1 per cent lactic acid, plus smaller amounts of other organic acids.[1] It is bacteriostatic and fungistatic, owing [2] to its pH of 5.73 to 6.19 and its content of acetic, propionic, caproic, caprylic and lactic acids.[3] Cosmetics must be designed so as not to destroy the natural defenses of the skin by completely neutralizing these acids for long periods.

The sebum is mostly fatty material, resulting from the disintegration of the superficial cells of the sebaceous glands. The oily coating normally present on the skin is instrumental in keeping it supple and moist; when it is lacking, cosmetics should supply such a coating. Since the rate of absorption into the skin depends [4] on the solubility of the material in the liquid filling the ducts of either the sweat glands or the sebaceous glands, cosmetics intended to be so absorbed must be designed to have such solubility.

Permeability of Skin. The question of permeability is of basic importance in the formulation of cosmetics. For make-up items which should beautify or protect the skin while covering it, but leave it unaffected after removal, absorption into the skin should be avoided. For treatment items, designed to improve the texture of the skin, absorption may be desired. We should distinguish clearly between absorption into the skin or into the immediate subcutaneous tissue, and absorption through the skin into the circulatory systems. The latter type of absorption, if it outweighed the former type, would cause a preparation to cross the border between cosmetics and pharmaceutical preparations. The cosmetic preparation exerts its effect in the area where it is applied, not indirectly by absorption into the circulation and redistribution into the tissues. Studies of permeability which depend on determination of a substance in body fluids, after application to the skin, are therefore of more interest for pharmacological than for cosmetic application. The most reliable method of studying take-up of substances by the skin tissues is probably the histological examination of the excised skin, which of course presents complications.

The human skin protects the body, not by totally preventing diffusion from outside into the tissues, but by slowing up such diffusion. Materials may penetrate the skin, and later be excreted mainly through the skin, without ever entering the blood to any significant extent.[4] The transfer of minute amounts of some materials in the skin has been followed by immunological reactions. Substances can penetrate the skin by three main

channels: (1) the sweat glands, (2) the hair follicles and the sebaceous glands, and (3) the keratin matrix. There has been but little study, and none of it quantitative, of the permeability of the horny keratin, but it should allow passage of gases and other small molecules. Water probably does not penetrate the keratin, although it does cause it to swell.

Sweat glands are lined with cells which, as part of their regular function, permit the passage of an aqueous solution, sweat, through their membranes. Under the proper conditions, it is reasonable to expect that these cells will absorb certain solutes from solutions which enter the sweat ducts from the skin surface; this absorption will not, however, be a simple physical diffusion, but will be governed by the physiological needs of the cell (see below). So penetration into the sweat pores will be the first step in the absorption of the major part of any water-soluble material. Similarly, the sebaceous glands will afford a gateway for penetration of fat-soluble substances.

Penetration into Pores. A liquid placed upon the skin has little tendency to enter the pores, for the air in the narrow channels is difficult to displace from outside. However, if the pores are filled with sweat or with sebum, as during rapid perspiration, or if the skin is covered with ointment or cream which closes the pores, the secreted liquid will gradually displace the air outward and fill the pores, thus forming a liquid bridge for diffusion of solutes from the ointment into the pores.

Abramson and Gorin [4] demonstrated that such penetration of pores by methylene blue is accelerated by electrophoresis. For transport by diffusion only, they give the following equation for the concentration, C, at any point in a cylindrical pore:

$$C = C_0 \left(1 - \frac{2}{\sqrt{\pi}} \int_0^y e^{-y^2} dy \right) \qquad \text{where } y = \frac{x}{2\sqrt{Dt}}$$

and C_0 is the concentration of the substance at the mouth of the pore, x is the distance from the mouth, D is the diffusion constant, and t is the time. In their experiments, methylene blue was retained largely in the cells lining the pores; histamine diffused into larger areas of the skin, as demonstrated by whealing. Histamine injected intracutaneously does not remain long in the skin, but when introduced through the pores it has been shown to remain over a week.[5] Eller and Wolff [6] found that fats penetrate the skin largely along hair shafts and in the ducts of oil glands. Animal fats showed the greatest depth of penetration, and low melting point facilitated penetration. Moore, Lamarre, and Beck [7] reported that the absorption rate of testicular hormone varied with different oils used as solvent, and with the concentration; sesame oil was especially effective.

Absorption by Cells. After a substance has an effective concentration

outside the cells lining the ducts of the glands, it must penetrate into these cells, then spread further into other cells, if it is to have any general effect on the subsurface tissues of the area. We can visualize such spreading as proceeding in at least four different ways: (1) directly, from one cell to another; (2) through the intercellular fluid; (3) through the lymph; (4) through the blood. It is very probable that more than one of these paths are involved in most cases.

Zweifach[8] remarks that the type of permeability shown by the capillary membrane can be explained without taking into account the cellular components of the vessel. The filtration barrier can be divided into two constituents, the intercellular cement (a reversible calcium salt of a weak acid) and an adsorbed layer of protein, which can be affected independently or concurrently by changes in the fluids bathing either side of the capillary. The physical state of the cell membranes can be varied by changing the pH or calcium content of the medium, so variations in permeability can be induced by conditions affecting the intercellular cements. We are not primarily concerned, however, with absorption of substances into the blood. Absorption into cells of the type forming the deeper layers of the skin, and the subcutaneous tissue, is our first concern.

We may consider, with Mazia,[9] that the ionic content of the cell is divided into two parts: the part at the surface, containing most of the calcium, is capable of rapid exchange with ions in the surrounding fluid; the part inside the cell contains most of the other electrolytes, and their concentration is relatively independent of short-run environmental conditions, being increased or decreased by true accumulation involving the *metabolic activity* of the cell. Biophysical investigations give evidence that the plasm membrane is not many molecules thick, containing lipid as major component. But, as Just [10] notes, direct biological investigation of the cells in action indicates the existence of a layer of microscopic thickness, the *cortex,* which is important in the functioning of the cell. Substances which enter the pores of the skin, but which cannot be used in the metabolism of the cell, may be passed along into the blood by transfer through or along this cortex, and therefore have no effect on the local tissues.

An analogy for the cell's behavior is found in the ability of proteins to bind heavy metal cations in relatively weak salts. Stearate films can exchange their metal ions for traces of copper in distilled water.[12] Assuming a molecular area of $20Å^2$, the number of molecules that could form a single layer on the surface of the cell is of the same order of magnitude as the number of molecules of calcium that can participate in rapid exchanges between cell and environment.[9] So a membrane of relatively few monolayers of lipid could account for the changes involving calcium

observed by Harvey and Danielli [13] in their study of yeast, *Elodea,* and *Arbacia* egg cells.

Mazia proposes, in extension of Brook's theory,[14] that the first step in accumulation of an ion is its binding at the cell surface, which through oxidative metabolic processes continually releases these ions into the cell interior and binds more ions from the medium. So in cells accumulating salts, the increase in electrolyte content should be found in the non-exchangeable fraction of the ions, the exchangeable or surface fraction remaining relatively constant. In yeast cells, increasing the calcium content of the medium results in increased calcium uptake, but the fraction that is rapidly exchangeable for potassium remains nearly constant. Blinks [15] also considers permeability and metabolism to be closely related cell functions, and believes that the latter may affect the former not only by setting up gradients of proteins, organic acids, etc., but also by influencing the nature of the cell surface.

Brooks [14] measured rates of diffusion of radioactively tagged alkali metal ions into living cells, and obtained rates of about 1.76×10^{-7} gm cm^{-2} hr^{-1}, instead of 10^{-8} gm cm^{-2} hr^{-1} estimated in older experiments. Both cation and anion move against a concentration gradient more rapidly than is explicable by any present hypothesis, but a mechanism of ion exchange must be involved, such that the cells can retain specifically-accumulated ions, notably K^+, during this rapid exchange, approximating the speed of free diffusion of these ions in water. When immersed in pure water, *Nitella* cells lose their salts slowly, compared with the rate of loss in solutions of other salts, for water supplies few ions for exchange; so ions do not appear to travel together through the cell surface, although the positive and negative ions pass at about the same rate. The passage of ions across the cell membrane may be visualized as occurring by a progression similar to that of individuals in a "Paul Jones" dance, made possible by a plasma membrane described as a mosaic.[16] Brooks [17] considers that the acidic and basic groups on proteins allow ion-exchange reactions similar to those in permutite. The plasma membrane appears to be a relatively dense structure, and practically all ions must be fixed close to the combining groups of the fats, proteins, etc. The amounts of radioactive ions taken up by *Nitella* were 20 to 1600 times the possible combining power of the plasma membrane materials, so most of them must have been contained inside the cell, not adsorbed on the surface.

When metabolism lags, the selective nature of ion intake is lost. Muscle cells, on contracting, lose K, and inert muscles may suffer a permanent loss of K, and atrophy. Brooks [14] asks whether ion content, some kind of metabolism, and vital activity are not an indissoluble complex. Hoagland [18] found that in the root cells of barley and wheat the internal conditions required for most rapid salt accumulation were high content

of sugar or other available carbohydrate and low salt concentration. The inward transport of salt against a gradient was considered to depend on aerobic metabolism, involving a process akin to secretion. During active salt absorption, the sugar content rapidly decreases, especially in the case of certain K salts. Steward [19] found striking effects on metabolic processes during active absorption of salt by potato tuber discs. Increased absorption of certain K salts caused increases in respiration and synthesis of protein from amino acids or other dissolved nitrogen compounds. Ca salts caused decreased respiration and protein synthesis, even though anion accumulation occurred. Steward concluded that K^+ is especially influential in stimulating metabolism during the process of entering the cell. These results may well lead us to believe that the proper combination of inorganic salts may be the most effective medium for influencing the metabolism of the tissues in and under the skin.

Desirable and Undesirable Attributes

The goal of cosmetic treatment can be described very simply; it is the attainment of a healthy, youthful appearance. Fashion may dictate the use of bright colors as makeup on lips and nails, and an individual's features may require the accentuation or minimizing of certain parts, but in general the user of "treatment" preparations would be very happy with the skin and hair of the normal twelve-year-old girl. The effect which is hoped for is largely the negation of the effects of ageing.

Changes Produced by Age. The process of ageing must be ultimately referable to the slowing of cell metabolism and division. In youth, the cells are continuously growing and dividing, and any non-diffusible waste products are automatically prevented from building up in concentration. When growth stops, the cells of the matrices producing skin, hair, and nails do not stop dividing, but their rate of division does lessen. Metabolic activity diminishes in the skin and subcutaneous tissues, as evidenced by loss of the pink flush so flattering to the youthful cheek and the appearance of wrinkles as the elasticity of the tissues decreases. It is not impossible that biochemical research may reveal a method of slowing up the ageing process, by control of diet or other general treatment; but the application of any such method would belong in the field of medicine, not cosmetics. However, the development of a local treatment which will increase the metabolism, promote healthy proliferation of skin, and remove undesirable metabolites, is entirely within the realm of cosmetics.

The hair changes markedly with age, owing to lessening activity of the matrix cells. One effect is loss of color, another is disappearance of the hair from part or all of the scalp. Neither effect has yet been successfully

prevented, either medically or cosmetically, but the loss of color can of course be compensated for by dyeing the hair.

Effects Required of Cosmetics

Cosmetics are expected to effect an alteration in one of the following attributes of skin, hair or nails: color, odor, shape, water or oil content. It may be interesting to consider separately these effects, and the agents used to produce them.

Alteration of Color of Skin. Cosmetics are generally designed to alter the color of the skin in the direction of the normal, youthful skin, although fashion may dictate some deviations from this goal. Face powders, which cover minor blemishes, absorb and cover oil to eliminate shine, and generally lighten the apparent skin color, are probably the most universally used agent for altering the color of facial skin. Greater adhesiveness is given to face powders by adding waxes, soaps or oils, converting them to cake make-up or make-up foundations, which cover the skin more heavily and lastingly.

To tint the cheeks, more highly colored rouges are used. These are either dry powders containing pigments in inert white base, usually compressed into a cake for convenient carrying; or pomades, with pigments dispersed in an oil-wax mixture. The latter type of rouge, cast in stick form, is widely used as lipstick, to color the lips. Here, the colors used are much brighter than the natural lip color, although all colors of lipstick which have found acceptance can be classed as reds. Liquid rouges have been promoted, both for cheeks and lips, but are not widely used.

The pigments used for face powders and cake make-up are largely inorganic substances such as iron oxides, sometimes brightened by addition of certified lakes of organic dyes. For lipstick and other rouges, the lakes of organic dyes form the bulk of the coloring matter. Lipsticks generally contain an indelible color in addition to the insoluble pigments, generally a halogenated fluorescein. A skin coloring preparation in vogue during the war, when hosiery was in short supply, was leg make-up; made mostly in sun-tan shades, this was based upon iron oxide pigments suspended in an emulsion containing film-forming gums, which dried on the skin to a more or less lasting colored film.

Sunscreens may also be considered as preparations for affecting skin color, although the other effects of sunburn are more unpleasant than the skin coloration. Henschke and Schulze [23] showed that erythema (sunburn) is produced by ultraviolet rays shorter than 3200Å, while pigmentation without erythema is produced by rays of 3200 to over 4000 wavelength. It is, therefore, possible to prepare sunscreens which will protect from both burn and tanning, or, as is more often desired, from burn only.

Many substances are available with strong absorption of rays below 3200Å, such as menthyl anthranilate, quinine salts, β-methylumbelliferone, etc., from which suntan oils, lotions and creams are made containing 2 to 8 per cent of the sunscreen. The darkening effect of the longer ultraviolet rays is attributed to their acceleration of the formation of melanin; while the injurious results of exposure to the shorter rays are reported to involve conversion of histidine to histamine, with consequent infiltration and discomfort.

For bleaching the skin, lotions or creams containing lemon juice or other weak acid were long used. Their action was at best very mild. Creams containing ammoniated mercury or other mercury salts are used to lighten the skin, and Nealon [24] reported that the use of a cream containing 3 per cent ammoniated mercury, applied one-half hour daily for six weeks, lightened the facial skin by 15 per cent of the observed white human range. From the work of Raper,[25] Arnow,[26] and Jones,[27] it appears that melanin may be formed in the skin from any of the three amino acids, tyrosine, phenylalanine or tryptophan. The formation from tyrosine involves oxidation to dihydroxyphenylalanine (dopa) catalyzed by tyrosinase or by ultraviolet light. The enzyme, dopa-oxidase, catalyzes the further oxidation to dopaquinone, which condenses to the bicyclic 5,6-dihydroxy-dihydroindole α-carboxylic acid. This compound, through the intermediate diketone, hallochrome, is converted to 5,6-dihydroxyindole-α-carboxylic acid, which by polymerization forms melanin. If the above sequence can be interrupted at any point, melanin production will be prevented and the color of the skin lightened. Mercury inhibits the copper-containing enzymes tyrosinase and dopaoxidase, and thus accomplishes the result. There are certain limitations to the use of mercury compounds on the skin, so the problem of controlling skin pigmentation at will is still not ideally solved.

Coloring of the Hair. Mascara and eybrow pencil are used to color the eyelashes and eyebrows, respectively. For these preparations organic dyes are not allowed, so mineral pigments and carbon black are used. The application of color to the hair of the scalp should be classed as a dyeing operation, rather than make-up. The process is not unlike that of dyeing fabric, and the same principles apply, except that hair is attached to the scalp and care must be taken not to injure the skin. This limits the usable dyes and the methods of application, and the operation is generally entrusted to a skilled beautician.

Evans [20] has summarized the development of hair dyes and their method of action on the hair. The early use of lead and silver salts, converted to corresponding sulfides on the hair, and of pyrogallol-metal salt mixtures, has given way to the phenylenediamine dyes. These, in spite of their tendency to cause allergic reactions in a small percentage

of persons, are the most popular hair coloring agents. They are sold in aqueous-alcoholic solution containing ammonia and soap or other wetting agent, and are mixed just before application with hydrogen peroxide which oxidizes the diamine to the complex dye molecule. The wetting agent helps the diamine and peroxide penetrate the hair shaft. The peroxide can open the disulfide cross-links of the keratin to give anionic groups. These groups may serve as points for attachment of the dye molecule; and the performance of the dye may be due to the fact that the large dye molecule, formed inside the lattice of the keratin molecule from smaller molecules which could penetrate the lattice, is itself too large to escape from the lattice.

Coloring of Nails. Nails are colored by the application of nitrocellulose lacquers. The formulation of nail lacquers is similar to that of other lacquers, but the pigments used must be harmless, certified colors, and the solvents chosen with more regard for pleasantness of odor. Rapid drying and high gloss are important. Adhesion to the nail may be improved by use of a tacky undercoat, but one variety, reportedly based on a synthetic rubber, caused allergic reactions in the nail bed and therefore must be avoided.

As reviewed by Wing,[21] and by Thomasset,[22] 1/2-second nitrocellulose is suitable for most lacquers and a typical composition would be 10 per cent nitrocellulose, 5 per cent of a plasticizer such as a phthalate or castor oil to improve flexibility and adhesiveness, 10 per cent of a sulfonamide-formaldehyde resin to improve luster, 3 per cent dyes and pigments, 5 per cent alcohol, 20 per cent ethyl acetate, 15 per cent butyl acetate, and 32 per cent toluene. The colors selected are certified reds fast to light, such as D & C Reds No. 11 and No. 34, and Ext. D & C Red No. 2, readily dispersible and nonstaining, plus purified iron oxides, iron blue and titanium dioxide. A viscosity of 270 to 310 centipoises is considered desirable.

Odorants and Deodorants. Although perfumes may in various epochs have been used to cover body odors and thus obviate the necessity of bathing, modern practice favors a clean skin as a prerequisite of attractiveness. Deodorants are designed to prevent unpleasant odors from developing on the skin, and perfumes, toilet waters, and the like are assigned a more positive role in creating a generally pleasant olfactory impression.

Klarmann [28] reviewed the cosmetic aspects of perspiration and its control, pointing out that the eccrine sweat glands found over most of the body secrete a clear fluid containing about 1 per cent total solids, of which sodium chloride constitutes an average of 0.7 per cent. Organic acids such as lactic, acetic, propionic, caprylic, caproic, citric and ascorbic form most of the remainder with traces of urea and uric acid. The apo-

crine sweat glands, usually connected with a hair follicle, occur mainly in
the armpits, on the chest, and in the genital areas, and since their secre-
tion involves a break in the tips of the secretory cells it contains nitro-
genous and fatty matter. The perspiration from these areas therefore
offers especially high potentiality for development of bad odor under
the influence of bacteria normally present on the skin.

Killian and Panzarella[29] demonstrated that perspiration passed through
a Berkefeld filter to remove bacteria, could be stored 24 hours without
change in odor or pH; unfiltered perspiration developed a strong, un-
pleasant odor in 24 hours, while its pH rose from 4.82 to 7.86 as urea
was converted to ammonia. Objectionable skin odors may therefore be
avoided by controlling perspiration or by preventing its bacterial[1]
decomposition.

Antiperspirants are generally based on salts of aluminum.* The alumi-
num forms albuminates with the proteins of the skin, and the tissues
around the mouth of the sweat duct are reported to swell, reducing the
aperture and the flow through it. *Simple deodorants* are based on anti-
septics and many have been used, such as oxyquinoline salts, sodium
perborate, zinc peroxide, etc. Combination deodorants and antiperspir-
ants contain compatible astringents such as aluminum chlorohydrate and
antiseptics such as hexachlorodihydroxydiphenylmethane. Cream deodor-
ants are the most popular, but sticks, powders and liquids are also sold.

Perfumes. Pleasant odors contribute immensely to the general impres-
sion created by any person. The effect of odor is largely irrational and
imponderable, but by no means negligible. Nearly all cosmetic prepara-
tions contain enough perfume to give some degree of pleasant odor after
application, although most of them are scented unobtrusively, so that
they will not interfere with any perfume which may simultaneously be
applied. Perfume materials were, until about 75 years ago, all of natural
origin; but synthetic products are now used in large amounts to fortify,
extend, enhance, and amplify the natural materials. There is little value
in a classification of perfume materials according to their chemical char-
acteristics, but the useful compounds include hydrocarbons, alcohols,
aldehydes, ketones, phenols, ethers, esters, lactones and nitro compounds.

Most of the natural materials are essential oils or resins obtained by
distillation, expression, extraction or maceration of the flowers, leaves,
twigs, wood, bark, roots, fruit, seeds or buds of some plant. A few
materials of animal origin, such as musk, civet, castoreum and ambergris,
are used as fixatives and possibly for their aphrodisiac qualities. From
these natural products there is a gradual transition to outright synthetics.
Improved essential oils, such as terpeneless oils, are altered by removal

* In his *Historia Naturalis,* the elder Pliny states (xxv. sec. 185) relative to alum: *Virus
alarum sudoresque sedat.* (It reduces offensive odors of the axilla and sweating).—*Ed.*

of the less valuable components, giving oils of higher value. *Isolates,* such as safrol from camphor oil, are separated from all other components of the oils containing them. *Modified oils,* such as vetiver acetate, are made by chemical treatment of oils to convert one or more components to a derivative more desirable for some purposes. *Synthetics made from isolates,* such as vanillin made by conversion of eugenol isolated from oil of cloves, are sometimes preferred to synthetics of other origin.

Synthetics, of chemical origin, such as vanillin made from guaiacol,* often can replace the natural product at much lower cost. *Novel synthetics,* such as amyl-cinnamic aldehyde supply materials not found in natural oils, but of value in simulating or extending natural materials. Others, such as the higher aliphatic aldehydes, are used to create popular perfumes not limited to imitation of any natural odor.

The blending of the many available materials into a perfume of desired type, quality and persistence is more art than science, for there are many gaps in our knowledge of the physical, chemical, physiological and psychological aspects of odor and olfaction. We know that odor is a powerful factor in creating moods and in determining attitudes toward either persons or objects, but our knowledge of odor-structure relations, and of odor-behavior relations, is largely empirical.

Waving of Hair. Hair is the only tissue which is subject to a definite change of shape by cosmetic treatment. Since wavy hair is generally considered to adorn the feminine head to a greater degree than hair of the straight variety, enormous effort has been expended on methods of waving hair, preferably in a permanent manner. It is estimated that something like one per cent of the average woman's adult life is consumed in the operation of hair waving.

Hair softened by wetting, wound spirally on a rod and so held until dry, will acquire a more or less lasting wave; this procedure is the basis of the use of curling papers, all sorts of patented curlers, and curling irons. Suter,[30] McDonough,[31] and Reed and associates[32] have summarized the development of mechanical and chemical methods of permanent waving. Early methods improved on the ancient home curling practices chiefly by using first borax, then ammonia, to speed and increase the softening of the hair, and by designing effective heating devices for speedy drying. Later, the cumbersome heater was replaced by pads heated by chemical reaction, using such combinations as $CaO+H_2O$; $Al+Cu\ SO_4$; $Al+$ various oxidizing agents; hexahydric alcohols $+$ permanganate.

More recently a cold process has been developed which utilizes ammonium thioglycolate to soften the hair, the disulfide bonds between chains being broken according to the equation:

* Vanillin is now synthesized in quantity from the crude calcium lignin isolated from waste sulfite paper mill liquors.—*Ed.*

Hair—S—S—*Hair* + 2HSCH$_2$—COONH$_4$ ⇌

$$NH_4OOCCH_2—S—S—CH_2COONH_4 + 2\textit{Hair}—SH$$

Since cross-links between chains are thus ruptured, the hair becomes more pliable and loses elasticity. Subsequently the thioglycolate is removed, with the hair still rolled in spirals, by rinsing with a solution of an oxidizing agent such as hydrogen peroxide or potassium bromate, which also reoxidizes the hair, restoring the disulfide linkage so that the hair, after drying, regains its elasticity, but with the new shape impressed upon the fiber. Other bonds beside the disulfide, such as peptide linkages, hydrogen bonds, and even van der Waal forces play a part in the process, but this part can at best be only minor, as shown by Reed [32] in experiments on stretching with monochloroacetic acid and with dilute HCl. In practice the pH of the thioglycolate must be carefully controlled: if too low, the solution is ineffective; if too high it is too irritating to the scalp and may injure the hair.

The foregoing processes can also be used to straighten excessively kinky hair. Any liquid mixture that softens the hair (loosens the cross-bonds) will facilitate change of shape under tension, and the hair will tend to retain whatever shape it is held in when the softening agent is removed and the fiber regains its stiffness. Furs are treated similarly to increase or decrease the curl, as desired, in order to simulate more valuable skins. Here stronger agents can be used, since the problem of irritation is removed.

Removal of Hair. The same chemical agents which soften hair will, if caused to react more strongly, destroy the keratin structure so that the hair disintegrates and can be washed off. Creams or solutions of metallic sulfides such as barium, strontium, or calcium sulfides will accomplish the purpose, but on account of the hydrogen sulfide evolved their use is attended by an unpleasant odor. Sulfide depilatories have therefore largely given way to those using ammonium thioglycolate, which have a less objectionable odor.

Stimulation of Hair Follicles. Removal of hair is much easier than stimulation of its growth; indeed, up to the present the stimulation of growth and pigmentation of the hair is a cosmetic ideal much striven for but not certainly attained. Both loss of hair and loss of pigment in the remaining hair are effects which increase with advancing age, and most attempts to overcome these effects have been based on methods designed to invigorate and stimulate the hair follicles. Hair tonics have been formulated with any of a long list of irritants which, used with vigorous massage, should increase the flow of blood to the scalp and so, hopefully, increase the activity of the hair follicles. Claims have been made for each of three members of the vitamin B complex as preventers of gray hair, baldness, or both; pantothenic acid,[33] *p*-aminobenzoic acid,[34] and in-

ositol.[35] These claims were based on animal experiments and have not
been confirmed by carefully controlled trial on humans, but a recent
patent[36] claims that cosmetic preparations containing salts of panto-
thenic acid can prevent falling hair, gray hair and dandruff.

There is some evidence that baldness is brought on by accumulation
of male sexual hormones,[37] but if so this effect has not been successfully
overcome, either by medical or cosmetic treatment. Both the adrenal
cortex and intake of sodium chloride seem to be involved in melanin
metabolism which is so important in hair pigmentation.[38] Traub[39] re-
cently summarized the present knowledge of hair growth and noted our
lack of fundamental information on many phases.

Improvement of Skin. It is probably the hope of every woman who
applies cream to her skin that she will thereby achieve a smoother,
softer, more youthful appearance. Certainly this effect is obtained to some
extent, by supplementing the oily material normally secreted by the
sebaceous glands, so that under conditions where the natural protection
is inadequate, the oils applied in the cream furnish protection, and pre-
vent excessive drying, chapping, and irritation. This protective effect is
transient, and cosmetic research has striven to find ways to stimulate cells
of the deeper layers of the skin, and possibly those of the underlying
tissues, so that the pink color and firm, smooth texture of youth will
return to a skin beginning to show the effects of age. One method which
seems to have had some success involves the use of estrogenic hormones.
It is well known[40] that these hormones can be absorbed *through* the
skin into the system, if applied in large quantities and in the proper
media, thereby exerting an effect bordering on pharmacology. By limiting
the concentration to about 10,000 I.U. per ounce, we have a true cos-
metic; for the therapeutic dose for systemic effect is generally 2000 to
10,000 units per single dose, by injection; as much as 50,000 units per
week are given in some instances. Klarmann[41] summarizes the evidence
that a cream containing 7500 to 10,000 units of estrogenic substance per
ounce does produce a beneficial local effect without systemic complica-
tions. The estrogenic hormones applied topically have been reported to
exert beneficial and stimulating effects on senile epidermal cells, correct-
ing their excessive dryness and scaling;[42] histological examination of
excised tissues showed regeneration of surface epithelium; and in the
mesoderm changes indicating greater water content, increase in number
and size of capillaries, and firmer bundles of collagenous fibers. Thus a
general effect of firmness and smoothness is obtained.

Other methods of making the skin look younger have been investi-
gated. As Eisele and Eichelberger[43] note, the skin contains considerably
more total available water than any other tissue of the body, and is one
of the most important water-storage organs. Perdigon[44] considers hydra-

tion of the skin the central problem of aesthetic dermatology, and that homeostasis of extra- and intracellular liquids is best preserved by an isotonic solution containing all the mineral elements of blood plasma, such as diluted seawater or Quinton's plasma. Quinton [45] showed a general analogy between the saline composition of sea water and that of blood plasma of invertebrates, and used this analogy as the basis of a therapeutic system for which he claimed remarkable esthetic effects throughout the body; many cases of eczema were rapidly cured, and psoriasis improved, owing to rehydration of epithelium. Perdigon [44] prepared oil-in-water creams on the base of sea-water diluted to isotonicity with blood serum, and claimed softening effects on dry, eczematous skin, with formation of healthy, supple epidermis. He reasoned that a curative factor is also a preventive, and that this cream should preserve skin and give the freshness, tonicity and elasticity which resist wrinkling and restore youthful appearance.

It is possible that alteration of the inorganic salt content of cells may affect their permeability to nutrients, and their metabolism, as already mentioned under cell permeability. To make a cell more active, we must increase its intake and consumption of nutrients. Little is known of the absorption or utilization of nutrients by cells of the cutaneous tissues, although Borghi [46] reported that arginine is oxidatively deaminated by skin slices. Puccinelli [47] determined that the amino acids do not stimulate regenerative processes in amphibian epithelia, but that hydroxyproline increases oxygen consumption of the regenerating skin owing to activation of connective tissue, which metabolizes it. The lack of vitamin A has been blamed [48] for dryness and scaling. Increased regeneration of epithelium followed application of cod-liver oil to the skin after wounds and burns.[49] Creams containing vitamins A and D were reported [50] to improve dry, flaking skins with crepelike wrinkles, restoring their normal pliability.

Biochemical Classifications of Cosmetics

Gattefosse [51] has classified cosmetics according to their effect on the biochemical factors controlling the state of the skin as follows:

Hydration Modifiers.
(1) Hydrating agents—glycerite of starch, glycerinated creams, water-rich creams, gum mucilage, glycerinated fruit juices, etc.
(2) Insulating agents, protective and screening—petroleum jelly, lanolin-type creams, cold creams, etc.
(3) Tanning agents—sun-tan lotions, alum and tannin preparations, astringents and deodorants.

Lipoid Content Modifiers.
   (1) Hydrophilic lipoids—lecithin creams, almond pastes, cholesteorol creams, etc.
   (2) Solvatizing agents—soaps, cleansing milks and creams, sulfated fatty alcohol creams.

Ionization Modifiers.
   (1) pH modifiers—acid creams, milks, and toilet waters, toilet vinegars, distilled floral waters, alkaline creams.
   (2) Oxidation-reduction potential (rH) modifiers—oxidizing agents such as oxygenated waters and creams made with terpenic oils. Reducing agents such as sulfur creams and lotions, and organic sulfur creams.

Pigmentation Modifiers.
   (1) Apparent modifiers—face powder, lipstick, rouge, nail enamel, hair dyes, sun-tan stains, etc.
   (2) Actual modifiers—melanodermia, etc.

Local Innervation Modifiers.
   (1) Soothing creams and milks.
   (2) Camphor-ice, water, milk, etc.

Modes of Preparation of Cosmetics

Cosmetics are prepared in consistencies ranking from liquids to dry powders, the particular presentation chosen depending upon effectiveness, ease of use, and esthetic appeal.

Solutions. Practically all perfumes and toilet waters are prepared as clear solutions, involving solution of the odorous ingredients, chilling and filtering. The preparation of perfumery materials from natural substances or from chemical intermediates, of course, involves extensive extraction, distillation, chemical treatment, etc. which is outside the scope of cosmetic chemistry. Other cosmetics of solution type are some nail polishes and removers, brilliantines, astringent lotions, liquid deodorants, liquid rouges, hair lotions.

Liquid Emulsions. Generally containing small amounts of oil, liquid emulsions are composed mostly of water and emulsifying agent (soap, gums, etc.). Hand and face lotions, beauty milks, and bleaching lotions are examples. Preparation involves thorough mixing of ingredients at an elevated temperature, in some cases by use of a homogenizer. Liquid powders are sometimes made as suspensions rather than emulsions, requiring shaking before use.

Creams. Emulsions with an oily phase of sufficient amount and high enough melting point to solidify are used for a variety of cleansing,

facial, massage, vanishing, deodorant and other creams. Their preparation involves heating the oil and the water phase separately, above the solidification point, slow mixing with thorough stirring, and fairly quick cooling to give a smooth texture and glossy surface. Much has been written on the science and technique of emulsions,[52] and hundreds of emulsifying agents are available.

Pomades. Oil-wax mixtures without water are used, some with solids in suspension, as rouges, deodorants, liquefying creams, make-up bases, etc. Their preparation involves grinding the solids in all or part of the base, melting the ingredients together and careful stirring, to avoid settling, until solidified.

Sticks. Stiffer pomades are cast into stick form. Lipsticks, deodorant sticks and eyebrow pencils are examples; these are prepared in the same way as the softer pomades, but poured into molds which are then chilled so that the sticks become hard enough to handle.

Cakes. Preparations consisting mostly of solid ingredients may be compressed or cast into hard cakes. These are designed for application either by rubbing off with a dry puff (compact rouge, compact face powder) or by rubbing up into a paste with a moistened puff or brush (cake make-up, mascara). The first type contain little binder and are usually mixed dry, moistened and pressed. The second type may contain soaps and waxes and may be either pressed or cast.

Powders. Preparations presented as dry powders include face powders, dental powders, talcum powders, foot powders.

Whatever the type, a cosmetic product must be of attractive color and consistency, enticingly perfumed and packaged, in addition to being formulated and compounded for effectiveness in the intended application.

REFERENCES

1. Schwarz, L., and Peck, S. M., "Cosmetics and Dermatitis," New York, Paul B. Hoeber, Inc., 1946.
2. Talbert, G. H., *Am. J. Physiol.,* 50, 433 (1919).
3. Peck, S. M., *et al., J. Investigative Dermatol.,* 1, 237 (1938); *Arch. Dermatol. Syphilol.,* 39, 126 (1939).
4. Abramson, H. A., and Gorin, M. H., *Cold Spring Harbor Symposia Quant. Biol.,* 8, 272 (1940).
5. ———, *Arch. Phys. Therapy,* 21, 261 (1940).
6. Eller, J. J., and Wolff, S., *Arch. Dermatol. Syphilol.,* 40; 900 (1939).
7. Moore, C. R., Lamarre, J. K., and Beck, N., *J. Am. Med. Assoc.,* 111, 11 (1938).
8. Zweifach, B. W., *Cold Spring Harbor Symposia Quant. Biol.,* 8, 216 (1940).
9. Mazia, D., *Cold Spring Harbor Symposia Quant. Biol.,* 8, 195 (1940).
10. Just, E. E., "Biology of the Cell Surface," Phila., Blakiston Co., 1939.
11. Northrop, J., and Kunitz, M., *J. Gen. Physiol.,* 7, 25 (1924).
12. Langmuir, I., and Schaefer, V., *J. Am. Chem. Soc.,* 59, 2400 (1937).
13. Harvey, E. N., and Danielli, J. F., *Biol. Rev.,* 13, 319 (1938).

14. Brooks, S. C., *Cold Spring Harbor Symposia. Quant. Biol.,* **8,** 171 (1940).
15. Blinks, L. R., *Cold Spring Harbor Symposia Quant. Biol.,* **8,** 204 (1940).
16. Höber, R., and Höber, J., *Arch. ges, Physiol.,* **219,** 260 (1928).
17. Brooks, S. C., *Protoplasma,* **8,** 389 (1929); *Trans. Faraday. Soc.,* **33,** 1002 (1937).
18. Hoagland, D. R., *Cold Spring Harbor Symposia Quant. Biol.,* **8,** 181 (1940).
19. Steward, F. C., *Trans. Faraday Soc.,* **33,** 1006 (1937).
20. Evans, R. L., *Proc. Sci. Section, T.G.A.,* **10,** 9 (Dec., 1948).
21. Wing, H. J., *Proc. Sci. Section, T.G.A.,* **8,** 9 (Dec., 1947).
22. Thomasset, P. A., *Proc. Sci. Section, T. G. A.,* **3,** 5 (June, 1945).
23. Henschke, U., and Schulze, R., *Strahlentherapie,* **64,** 14 (1939).
24. Nealon, D. F., *Proc. Sci. Section, T.G.A.,* **1,** 7 (Jan., 1944).
25. Raper, H. S., *Biochem. J.,* **20,** 735 (1926).
26. Arnow, L. E., *J. Biol. Chem.* **118,** 531 (1937).
27. Jones, K. K., *Rept. External Prods. Research Inst.,* Vol. 2, Dec., 1940.
28. Klarmann, E. G., *Am. Perfumer,* **51,** 33 (1948).
29. Killian, J. A., and Panzarella, F P., *Proc. Sci. Section, T.G.A.,* **7,** 3 (May, 1947).
30. Suter, M. J., *J. Soc. Cosmet. Chemists,* **1,** 103 (1948).
31. McDonough, E. G., *J. Soc. Cosmet. Chemists,* **1,** 183 (1948).
32. Reed, R. E., Den Beste, and Humoller, F. L., *J. Soc. Cosmet. Chemists,* **1,** 109 (1948).
33. Unna, K., *J. Nutrition,* **20,** 565 (1940).
34. Martin, G. J., and Ansbacher, S., *J. Biol. Chem.,* **138,** 441 (1941).
35. Woolley, D. W., *Science,* **92,** 384 (1940).
36. F. Hoffmann-LaRoche & Co., A.-G., Swiss Patent 245,907 (1947).
37. Hamilton, J. B., *Trans. N. Y. Acad. Sci.,* **8,** 101 (1946).
38. Spoor, H. J., and Ralli, E. P., *Endocrinology,* **35,** 325-37 (1944).
39. Traub, E. F., *J. Soc. Cosmet. Chemists,* **1,** 129 (1948).
40. Zondek, B., *Lancet,* **1,** 1107 (1938).
41. Klarmann, E. G., *Am. Perfumer,* **48,** 41 (Nov., 1946).
42. Goldzieher, M. A., *Geriatrics,* **1,** 226 (1946).
43. Eisele, C. W., and Eichelberger, L., *Proc. Soc. Exptl. Biol. Med.* **58,** 97 (1945).
44. Perdigon, E., *Soap Perfumery Cosmetic,* **21,** 462 (1948) [from *Ind. de la Parfumerie*].
45. Quinton, R., "L'eau de Mer Milieu Organique," Masson, 1904.
46. Borghi, B., *Klin, Wochschr.,* **18,** 507-8 (1939).
47. Puccinelli, V., *Sperimentale,* **92,** 580-6 (1938).
48. Frazier, C. N., and Hu, C. K., *Arch. Dermatol. Syphilol.,* **33,** 825 (1936).
49. Lohr, W., *J. Am. Med. Assoc.,* **102,** 1991 (1934); **103,** 954 (1934).
50. McLaughlin, R. R. M., *Med. Times,* **67,** 303 (1939).
51. Gattefosse, R. M., *Ind. de la Perfumerie* (Oct., 1947).
52. Clayton, W., "Theory of Emulsions," 4th ed., Phila., Blakiston Co., 1943.

THE RECLAIMING OF ELASTOMERS

D. S. le Beau

Midwest Rubber Reclaiming Co., East St. Louis, Ill.

Introduction

The invention of the vulcanization of rubber antedated the mass-productive rubber plantations in the Far East, and while it gave to the rubber industry the necessary foundation for future development and expansion, it also created a demand for rubber which could not be met at that time. This demand, in turn, created the need for a means of reclamation of the vulcanized rubber scrap. The literature shows that some of the earliest patents for the reclaiming of waste rubber date back to about 1850. The first type of rubber article made commercially was rubber footwear, and consequently this provided the source of the first large volume of vulcanized rubber scrap available for the production of reclaim. From about 1890 on, the bicycle industry expanded considerably, and during the first decade of the twentieth century the production of carriage tires was also large. Both of these items subsequently served as raw materials for the reclaiming industry. At present, about 85 per cent of all the rubber scrap reclaimed consists of automobile tires.

Consumption and Production

It is of interest to follow the consumption of new and reclaimed rubber in the United States over a period of years (Figure 1). These figures indicate not only the size of the reclaiming industry in this country, but also the proportions of reclaim versus crude rubber which have been used. Approximately 30 per cent of all the rubber used in this country is reclaim. Undoubtedly this was partly due to the outright economy achieved because of the price of reclaim. Figure 2, however, shows that even when the price of crude rubber was less than that of reclaim, the latter still maintained its volume usage. Therefore, the use of reclaim in rubber goods must offer other advantages than those attributable to price economy only.

From Ball, "Reclaimed Rubber" (1947)
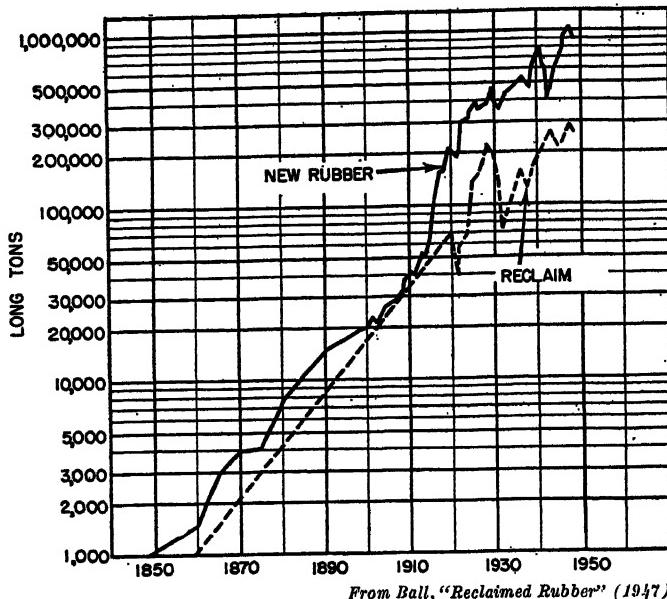
Figure 1. Consumption of new and reclaimed rubber.

From Ball, "Reclaimed Rubber" (1947)
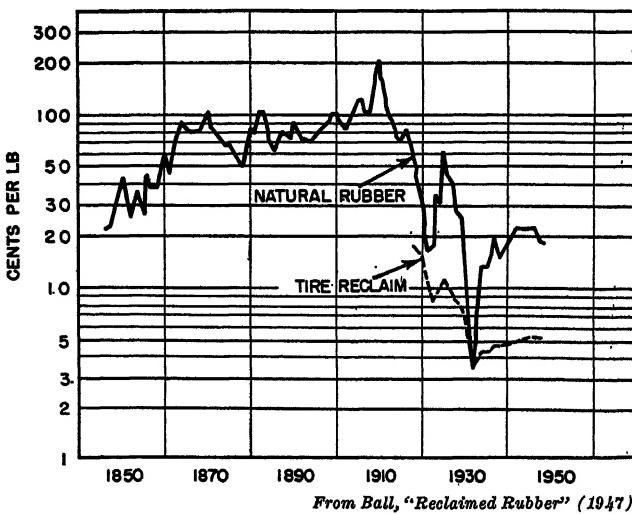
Figure 2. The price of natural (i.e., crude) rubber and tire reclaim.

The development of the various reclaiming processes was, in the beginning, closely related to the kind of rubber goods available and the compounding technique used in the production of these articles. The removal of fibers became necessary when quantities of bicycle tire scrap

were reclaimed, and two of the original reclaiming processes were initiated at that time. Accelerated vulcanization and the consequent change in the properties of the vulcanizates, the introduction of reinforcing carbon black into the rubber compounds, as well as the production and large-scale use of synthetic rubber, prompted the development of new reclaiming processes. Even within the last three years the progress made in compounding technique has changed the vulcanized scrap to a considerable extent. Thus, reclaiming processes have to be adapted to conform to continuously changing raw materials.

Devulcanization

The term devulcanization has been used quite indiscriminately to describe all kinds of reclaiming processes. Originally, vulcanization referred to the changes brought about by chemical addition of sulfur. By this definition, the reclaiming of vulcanized rubber cannot be called devulcanization because, regardless of the process, the removal of sulfur combined by primary valency forces is not achieved, even though claims have at times been made to the contrary. Whenever the combined sulfur was removed, the resulting product became useless. However, the vulcanization procedure has changed radically during the last ten years: polychloroprene is vulcanizable by the addition of metallic oxides; the amount of combined sulfur decreased considerably when organic accelerators were introduced, while at the same time the physical properties of the vulcanizates improved. Today the concept of sulfur vulcanization—insofar as it concerns the use of organic accelerators—is not purely chemical; the importance of secondary forces is also recognized, and the physical combination of sulfur is evaluated more as a means of bringing about network formation within the hydrocarbon polymers.

Calculations have shown that within the molecular-weight ranges commonly encountered in rubber, very few cross links have to be formed to insolubilize the polymer. Thus, the concept of chemical combination of sulfur is now considered in terms of chemical as well as physical forces. As will be discussed later, reclaiming is believed to be the *depolymerization* of the vulcanized polymer. Since vulcanization, in its present concept, can be considered polymerization of the polymer, the term *devulcanization,* as applied to the reclaiming process, does not appear as erroneous as it had seemed.

Major Reclaiming Processes

Principally a reclaiming process will have to depend on the application of energy, whether in the direct form of heat, mechanical working creating

heat infrared irradiation, electronic heating, or bombardment by atomic fragments. At present, energy is applied by direct heat, since this is more economical today than other forms of energy. We can distinguish between three major lines of reclaiming processes: (1) mechanical processes; (2) solvent processes; and (3) processes making use of high temperatures and reclaiming catalysts in the presence of steam or water.

Mechanical Processes. In these processes, which were developed at an early date, a large quantity of softener is added to the vulcanized scrap while the latter is worked on rolls. The reclaim produced is usually of inferior quality. While there may be some breakdown of the molecular chains, the large quantity of plasticizer used penetrates the mass and acts as a physical plasticizing agent, pushing the polymer hydrocarbon chains apart and changing the forces acting between the chains. This process is not used to any great extent, partly owing to the nature of the process and partly because of the fiber which is present in the major part of all the scrap.

If, however, mechanical working is carried out at high temperatures (200°C or over), and oxygen is kept at a low level, reclaiming of the scrap can be achieved by breakdown of the polymer chains.* The addition of small quantities of reclaiming catalysts increases the rate of the process. The process can be made continuous if a screwtype plasticizer apparatus is used to carry out the mechanical working, rather than the conventional two-roll plasticizer. Whether high or low temperatures are used, the presence of traces of oxygen is necessary.

Solvent Process. The solvent process is employed mostly in Europe for natural rubber scrap, and relies on the solubilization of the vulcanized waste by means of elevated temperatures and pressures in the presence of large quantities of solvents. Actual solution of the vulcanized rubber does not occur, but rather a dispersion of it in the solvent. The dispersed particles are of colloidal size and are of a more corpuscular shape than those present in crude rubber, because chemical linkages between the chains of the vulcanized polymer have not been broken. Viscosities of such solutions are quite low in comparison to those obtained from crude rubber. Yet the dispersion of rubber produced in this way is rather fine and textile fibers which may be present are removed by filtration. The process is not used in the United States as it is considered uneconomical because it calls for large quantities of solvents and adsorptive towers for the collection of the latter. On the other hand, the cellulose fiber which is obtained on filtration is clean and not highly degraded and can find multiple usage.

The reclaim remaining after solvent evaporation appears to be highly depolymerized and in that state is unsuitable for use. Simple mechanical

* See U. S. Patents 2,221,490 (1940); 2,408,296 (1946); 2,461,192 (1949); 2,461,193 (1949).

working at elevated temperature (about 315°F) seems to cause repolymerization of this reclaim, as evidenced by a toughening of the mass. However, if such working is carried out too long, a powdery brittle product is obtained which has no resemblance to rubber.

Processes Making Use of High Temperatures and Reclaiming Catalysts in the Presence of Steam or Water. These reclaiming processes are most generally practiced today, mainly owing to economic factors which are in turn conditioned by the kind of rubber scrap available in large volume. Essentially, the methods rely on the action of heat, in conjunction with steam and small amounts of reclaiming catalysts. Fiber, which is usually present (tire scrap), can be mechanically removed prior to reclaiming, or the two processes may occur simultaneously. In the latter case, the cellulose is destroyed by conventional methods of cellulose hydrolysis. Reclaiming is carried out either in a single-shell autoclave (*heater, pan process*) in direct contact with steam, or in a jacketed autoclave (*digester*), the vulcanized rubber scrap being dispersed in an electrolyte solution. A combination of both methods is possible, using a single-shell autoclave and introducing steam directly into the suspension of vulcanized rubber and electrolyte contained therein.

After completion of the reaction, the scrap is washed free of cellulose decomposition products, pressed, dried, and subjected to sheeting between rollers with a high friction ratio (*refiners*). During this stage, any nondevulcanized particles fall off as *tailings*. The sheets obtained are approximately the thickness of paper and are wound up on a drum where, because of their coherence and tack, they form a solid slab. This process is used to such an extent and presents so many colloid chemical problems that it warrants detailed discussion.

Effect of Particle Size and Shape on Reclaiming

While the size of the scrap particles is undoubtedly above the colloidal range, the effect of the particle surface is of predominant influence in reclaiming.

The tire scrap is first cracked and ground between corrugated rollers, conveyed onto screens, and sifted. The oversize particles are returned for further grinding. The ground scrap contains the cotton and rayon fibers mixed with the rubber and partly adhering to it. If natural rubber only is present, the size of the particles will not be as important as in the case of the synthetic product. Natural rubber depolymerizes quite readily upon application of heat, sometimes even in the absence of reclaiming oils and catalysts, and becomes plastic again. Of course, it can be expected that this depolymerization will be less uniform if the particles become too large, with the result that highly depolymerized

parts of the mass may still contain some fairly hard particles. It is well known that materials of widely different plasticity will not mix satisfactorily. Therefore, during the refining operation the less plastic particles will be ejected from the refiners in the form of tailings.

During grinding, the size of the particles are of extraordinary importance when synthetic rubber only is present, because of its peculiar property of heat hardening (repolymerization) which is predominant over its depolymerization. For a mixture of both polymers, and in the presence of cotton or rayon, other factors complicate the grinding process. The fibers do not grind at the same rate as the rubber polymers. Moreover, GR-S polymer has a tendency to grind faster than natural rubber polymer. This may, in part, be explained on a mechanical basis: natural rubber generally shows greater adhesion to the fibers than the synthetic polymer, the fiber usually grinds slower than the polymers and so any polymer particles adhering to it, and possibly of adequate size to pass the sizing screen, would still be returned to the grinders because the composite of fiber and polymer would not pass the sizing screen. In part, however, the difference in the rate of grinding between the two polymers may also be attributed to their inherent properties. Grinding with present-day equipment involves a cutting as well as a tearing action; natural rubber, because of its molecular-weight distribution, and the size and configuration of its molecules, is in general more resistant to this action.

If reclaiming oils and catalysts are added in the process, as is customary, the conditions of comminution become even more important, because the quantity of the above substances added to the total ground scrap is usually small, within a range of 0.5 and 20 per cent of the weight of scrap. These reclaiming catalysts and oils are expected to penetrate the polymer particles. Two phenomena of major influence will have to be considered if the scrap consists of both natural and synthetic rubber.

(1) The rate of diffusion of oils into the scrap particles decreases as particle size increases. The amount of oil present at any given point within the particle will vary with the distance of this point from the surface of the particle.

(2) The speed of diffusion of oils for natural and synthetic polymers is not the same. While the speed of oil uptake will also depend on the compounding and on the state of cure of the particular scrap, natural rubber in general permits a quicker diffusion of oils than GR-S.

Considering the effect of the particle size on the rate of diffusion, it would appear desirable to decrease the particle size as much as possible, to insure a more uniform penetration and prevent hard particle centers. This, however, causes a tremendous increase in the total surface of the particles. To obtain a halfway uniform distribution of reclaiming oil

and catalyst under such conditions would necessitate the wetting of at least part of the surface of each particle. While reclaiming oils are usually hydrophobic in nature and therefore will wet the polymer particles preferentially, their volume is so small compared to the tremendous total surface of the particles which are to be wetted, that a thin film of oil covering each particle will not be obtainable. The facts are brought out in the extreme by Table 1, where the effect of the particle size of the scrap

TABLE 1. EFFECT OF GRINDING SYNTHETIC TIRES
AND PEELINGS*

	Scrap Mesh Size	
Production	(ground to pass 5 mesh)	(ground to pass 1 mesh)
Pounds reclaim produced	30,041	29,112
Refiner hours	144	202.5
Pounds per refiner hour	208	144
Per cent tailings	8.7	17.2

* Kilbourne, F. L., *Rubber Age* (February, 1949).

on the production of reclaim is measured as refiner output. This table shows clearly that particles of large size will decrease the efficiency of the reclaiming operation, not only in machine output, but also in the amounts of tailings produced. (The experiment where the scrap was ground to pass a one-inch hole was an exaggeration; particles of this size do not occur under ordinary reclaiming conditions.) The increase in tailings produced from large particles is undoubtedly due, in part, to uneven penetration of reclaiming oil and catalyst.

The difference in the rate of oil diffusion between natural rubber and synthetic rubber complicates the situation further. Thus, natural rubber (i.e., that part of the polymer mixture which, as we shall see later, is more amenable to reclaiming) will obtain a larger share of the reclaiming oils whereas GR-S suffers from oil and catalyst starvation. This condition will, of course, tend to produce a soft mass from that part of the scrap which is natural rubber, while the portion which is GR-S will not plasticize as much and may even harden. The consequence will be a non-uniform mass, producing a considerable quantity of tailings.

Considering the points raised in regard to surface area, speed, and rate of diffusion of oils, it would seem advantageous to obtain oblong rather than corpuscular particles. An oblong particle shape would permit a minimum amount of comminution, while resulting in increased surface area for a given particle volume. The effect of rate of diffusion would also be decreased. The ideal particle shape would be that of a platelet or a sliver. While it is possible to change the openings in the sizing screen to an oblong shape, the majority of the particles ground with present equipment, and this holds particularly true for GR-S scrap particles, have more of a corpuscular than a plate-like or sliver shape. Nevertheless,

some improvement has been obtained from sizing screens with oblong openings.

Influence of the Surrounding Medium on Reclaiming

Inasmuch as the purpose of reclaiming is to convert the highly elastic vulcanized scrap to a plastic workable state, we can expect that the best means of accomplishing this is to cause molecular breakdown of the polymer chains with minimal chemical changes. This would result in a highly polydisperse system of generally shorter molecular chain length. However, it should be remembered that both natural rubber and GR-S are very sensitive to the presence of even small quantities of oxygen. This sensitivity increases with rise of temperature. Natural rubber absorbs more oxygen at any given pressure than GR-S; thus it can be expected that the vulcanized scrap will contain some oxygen absorbed before reclaiming is begun. The quantity of oxygen will depend on the conditions under which the rubber was used, on the polymer itself, on the compound, and to some extent, on the state of cure. Unvulcanized natural rubber, if exposed to elevated temperatures in the presence of oxygen, shows a progressive decrease of its gel component. While the influence of the thickness and surface area (particle size) of the sample subjected to oxidation was hardly noticeable at lower temperatures, its effect became very pronounced above 70°C. Also, depending on the dimensions of the particles in the sample, a critical temperature was found below which enough oxygen dissolves in the polymer so as to permit a uniform and slow attack. At higher temperatures, however, the rate of attack of oxygen is so greatly increased that mainly the surface of the polymer is affected.

In the case of GR-S, diffusion does not control the rate of reaction, provided that the thickness of the sample stays below 0.08 inch at 100°C or below 0.040 inch at 120°C. The marked influence of temperature on the reaction is again discernible. The state of cure of GR-S seems to have less influence on the reaction than that of a natural rubber compound. However, the presence of reinforcing carbon black strongly affects the oxidation of the polymer, and while comparatively little information is available for natural rubber, a series of accurate experiments on GR-S carbon-black compounds have yielded interesting results. Not only the quantity of carbon black, but also the temperature, exerts considerable influence on the extent of oxygen absorption. Furthermore, channel black was almost twice as effective as furnace black in promoting oxidation. In the initial stages of oxidation the presence of carbon black was found necessary for the activation of oxidation centers, the number of which increased with increasing carbon-black load.

The absorption of oxygen has been found to lower the physical properties of the polymer considerably. For the purpose of reclaiming it is advisable to decrease oxygen absorption as much as possible and to prevent nonuniform oxygen attack. Therefore, it has been found desirable to carry out reclaiming in an oxygen-diluted atmosphere at
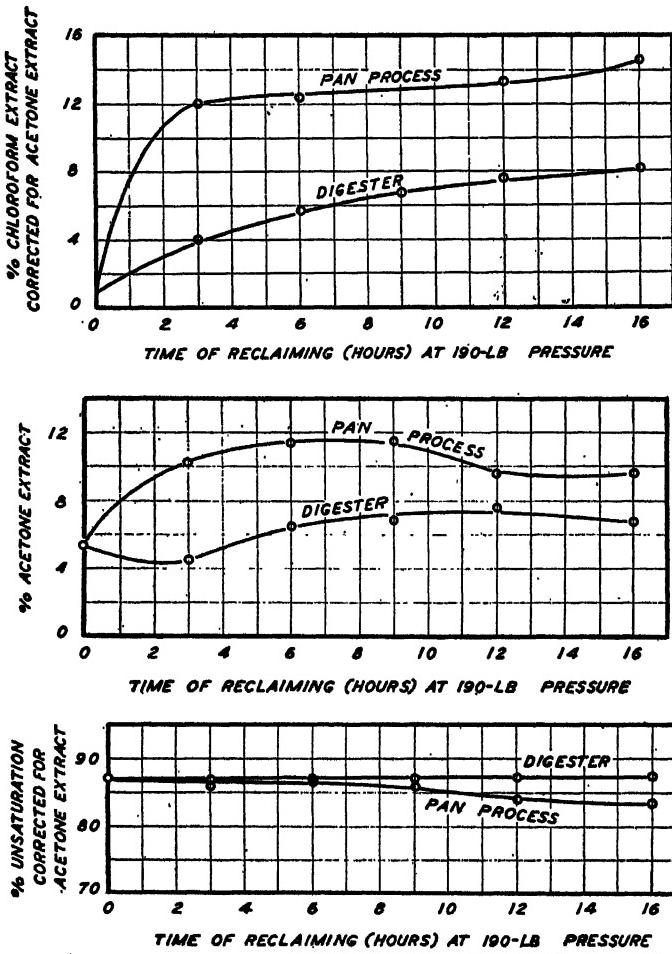
From le Beau, India Rubber World (April, 1948)

Figures 3, 4 and 5. Influence of the surrounding medium on reclaiming.

elevated temperatures. This condition is essentially fulfilled when reclaiming is carried out at high temperatures in the presence of steam or water. Because of the great sensitivity of the polymers to oxygen, the influence of the surrounding atmosphere even under such conditions must be considered important. Usually temperatures between 150 and 250°C are employed, and reclaiming which is carried out in open steam

(pan or heater process) proceeds at a greater rate than that which is carried out in water (digester process) of the same temperature. The rate and the quantity of chloroform-soluble polymer fraction which is produced during reclaiming are usually considered as a measure of the molecular breakdown. At present, very few data are available which disclose the course of the reclaiming reactions of GR-S. The following discussion is limited to natural rubber, and the data refer to the changes which occur in this polymer during reclamation in the absence of any oils or catalysts.

Figure 3 compares the quantity and rate of chloroform-extract formation for the heater and digester processes. The heater or pan process produces a greater amount of chloroform-soluble polymer fragments, within a shorter time, than the digester process. The quantity of acetone-extractable material formed during reclaiming (Figure 4), which can be associated with the formation of oxidized matter, is also greater if reclaiming has taken place in open steam and the unsaturation (Figure 5)

TIME OF PAN PROCESS RECLAIMING (HOURS) AT 190-LB PRESSURE
From le Beau, India Rubber World (April, 1948)
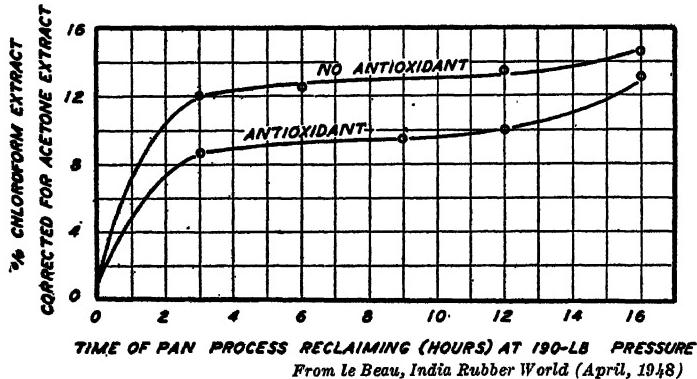
Figure 6. Influence of the presence of antioxidant on reclaiming.

of such a reclaim decreases slightly with the time. This would seem to emphasize that even small changes in the quantity of oxygen available during the process exert a marked influence. Steam contains about 5 per cent of oxygen, the ground vulcanized polymer will contain some oxygen and fresh oxygen is permitted to enter whenever it is necessary to maintain the pressure. If, however, the polymer is reclaimed while dispersed in water, we can expect that less oxygen is available. Actually, the rate of molecular breakdown is smaller under these conditions and oxygen attack does not appear to any pronounced degree (Figure 4). It has also been known for a number of years that, if oxygen was introduced purposely into the digester during the reclaiming period, the rate of breakdown was enhanced and reclaiming was speeded up. In contrast thereto, a rubber compound containing antioxidant will show less molecular break-

down if reclaimed in either open steam or water (Figure 6) and the acetone extract formed during reclaiming is also decreased.

The chloroform extract does not consist of "unvulcanized" polymer. Combined sulfur was found therein and determined by analysis. Therefore, the polymer chains present in the chloroform extract will contain cross-linkages. Since the introduction of even a very small fraction of cross-linkage has a tremendous influence on the solubility of a polymer, we must assume that the part of the polymer which dissolves in chloroform is of rather low molecular weight. Furthermore, polymer particles contained in the chloroform extract must be of a more corpuscular shape than those in the crude polymer.

Influence of Chemical Defibering Agents on Reclaiming

Because most of the vulcanized polymer scrap contains fiber (from tires), some means for the removal of the latter had to be worked out.

From le Beau, India Rubber World (April, 1948)
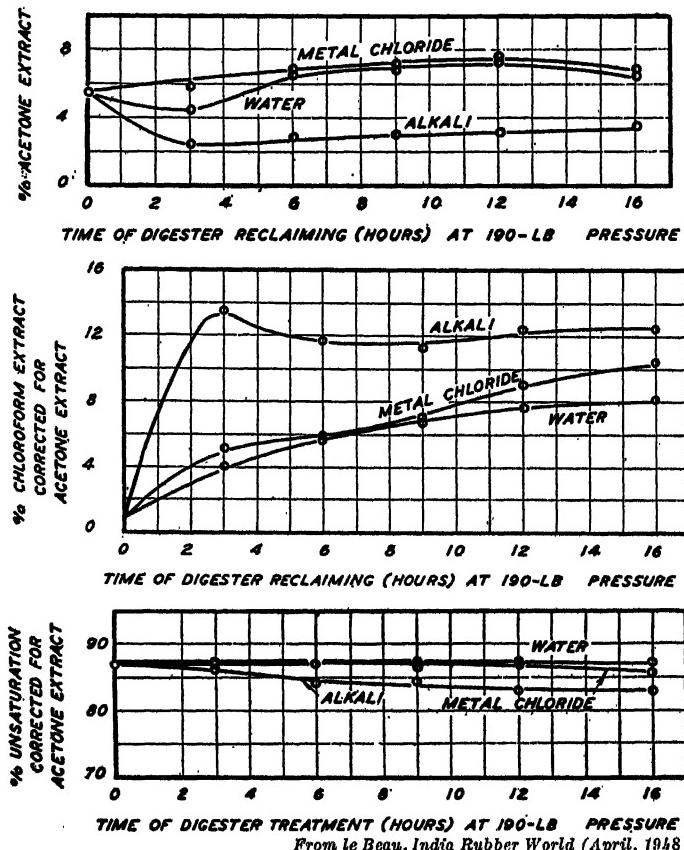
Figures 7, 8 and 9. Influence of chemical defibering agents on reclaiming.

Although heating in water or steam to high temperatures for a sufficient time will result in destruction of the fiber, this process is rather slow. Therefore, fiber-hydrolizing agents have been used early in the development of the reclaiming technique. Solutions of acids, alkalies and metallic chlorides, e.g., zinc or calcium chlorides, serve this purpose. The fiber decomposition products are removed by washing. The actual cellulose decomposition reactions are not sufficiently well known. There is no chemical equivalence in the effect between caustic or metallic chloride defibering agents. However, these defibering agents also influence the reactions occurring in both natural rubber and GR-S reclaiming processes. It was found that the acetone-extractable material does not increase to a large extent
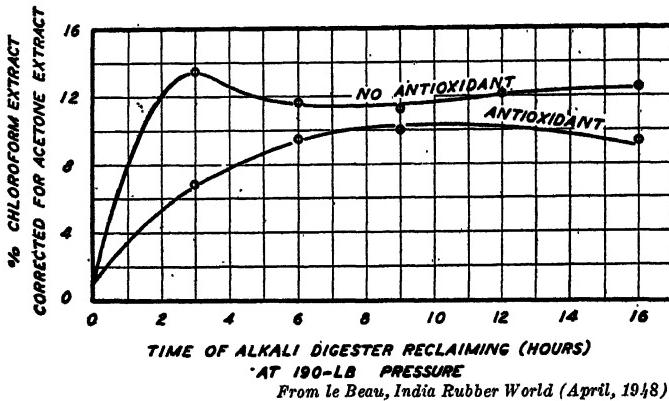
Figure 10. Influence of the presence of antioxidant on reclaiming.

during natural rubber reclaiming in the presence of a metallic chloride (Figure 7); the quantity of acetone-extractable material remains almost identical with that which has been observed for reclaiming in water. In the presence of alkali some of the saponifiable matter originating either from the rubber itself, or from added compounding materials, dissolves during the washing of the reclaim, because of the formation of soluble soaps. This reaction causes an apparent decrease in the acetone extract. However, the amount and rate of molecular breakdown of the natural rubber polymer depend on whether the reclaim was prepared in an alkaline or acid medium (metallic chloride). It should be pointed out that reclaiming carried out in the digester in the presence of water only, will result in slight, but discernible, acidity of the reclaim. This phenomenon explains why the data obtained for such reclaim and for that prepared in the presence of metallic chloride appear almost identical. It also emphasizes the importance of the acidity or alkalinity of the surrounding medium in reclaiming. Both the rate and amount of molecular breakdown were found to be much greater in the alkaline medium (Figure 8). A slight decrease in unsatura-

tion (Figure 9) could be noticed for reclaim prepared in an alkaline medium. The chloroform extracts obtained from reclaims prepared in an alkaline medium were found to have unsaturation values identical with those obtained from the original compound, and also exhibited the phenomenon of "recovery" (toughening) during storage. Chloroform extracts obtained from reclaims prepared in an acid medium showed a slight decrease in unsaturation.

If antioxidant was contained in the natural rubber scrap the rate of molecular breakdown in alkaline reclaiming medium was decreased (Figure 10); however, no change in acetone-extractable material was observed. If vulcanized natural rubber scrap which has been subjected to air ageing is reclaimed, the acetone extract of the resulting product increases. Both the rate and amount of molecular breakdown increase also, regardless of whether reclaiming was carried out in open steam, water, alkali, or metallic chloride solutions, whereas the unsaturation does not change. (Conditions of ageing were not so drastic that the unsaturation of the scrap itself changed.)

Oxidation of Rubber during Reclaiming

Since the marked effect of oxygen on the polymer was known, it was of interest to determine whether any permanent chemical linkage took place between the natural rubber polymer and oxygen during the reclaiming process. Unsaturation analysis had shown little, if any, change to have occurred during reasonable time intervals either in the reclaim or in that part of it which had become chloroform-soluble. However, it was thought possible that this analysis might not reflect the permanent introduction of very small quantities of oxygen. Infrared spectrographical analysis of the reclaims showed that $C=O$ and COH groups could not be found in water- or metallic chloride-digested reclaims. Pan-process reclaims and those prepared in an alkaline solution showed some absorption in the regions of the wavelengths associated with the above two groups. However, this absorption was less than that observed for the vulcanized scrap which had not yet been subjected to reclaiming.

Unfortunately, the various vibrational possibilities of ether linkages ($C—O—C$) have not yet been fully explored. These linkages could occur either inter or intramolecularly. However, they could not appear to any extent at the double bond, because the unsaturation of the polymer remained practically unchanged during reclamation. If, on the other hand, linkages between polymer chains occurred at the α-carbon atom, it can be expected that the quantitative chromic-acid oxidation of the natural rubber polymer would be affected by it. Data available at this time show that unless reclaiming is carried out under highly exaggerated oxygenating

conditions, the quantitative chromic-acid oxidation is very close to that obtained from the vulcanized, but not reclaimed, scrap.

Storage of Reclaim

The reactions occurring during the reclaiming process seem to continue during storage of the reclaim. The extent to which this will occur depends on the compound of which the scrap consists and on the time of reclaiming. Usually natural rubber reclaim prepared from highly reinforced compounds will show a decrease in plasticity and chloroform extract during storage. This is very similar to the well-known recovery phenomenon observed in unvulcanized rubber compounds after milling. However, natural rubber reclaims prepared from vulcanized compounds which did not contain any fillers prove to be rather unstable during storage; the shorter the reclaiming time the less stable the reclaims appear to be. Whereas the reclaims prepared in water or metallic-chloride solution do not exhibit a great change in their acetone extracts, the latter increases considerably during storage if the reclaim was prepared in the presence of an alkaline solution. Similar results are obtained when measuring the molecular breakdown, showing that during storage the alkali reclaim increases considerably in chloroform extract, particularly if the reclaiming period had been short; whereas the reclaims prepared in the presence of water or metallic chloride change only moderately. The unsaturation of the alkali and pan-process reclaims showed some decrease. In contrast thereto, the unsaturation of the reclaim prepared in water or in a metallic-chloride solution hardly changed at all. These results show that the reclaims prepared in the presence of alkali or by the pan process are less stable than those prepared in acid environment, i.e., in the presence of water or metallic-chloride solution.

Reactions Occurring During the Reclaiming of Natural Rubber

A discussion of the reactions occurring during the reclaiming of natural rubber must include our present knowledge of the steps in the molecular breakdown of the crude polymer. It is well-established that rubber can be heated up to 200°C without any molecular breakdown, provided that even traces of oxygen are absent. On the other hand, heating natural rubber to a temperature far below 200°C, in an atmosphere containing small quantities of oxygen, will cause its molecular breakdown. Milling natural rubber in the absence of oxygen will result in some plasticity increase; if traces of oxygen are present, however, considerable plasticity increase is obtained without any change in the unsaturation of the polymer (within reasonable time limits of milling). The plasticity decreases again if the rubber is stored after milling. The formation of hydroperoxides has been

proven to occur during milling. Hydroperoxidic reactions following a chain mechanism have been encountered in olefins and they have been discussed with special references to such long-chain olefins as natural rubber.

Vulcanized natural rubber can be reclaimed at elevated temperatures by mechanical working, provided that traces of oxygen are present. Actual permanent combination of oxygen with the reclaim has so far not been found. The introduction of oxygen into the digester vessel will speed up reclamation in the presence of steam or water. A small amount of oxygen is present at all times during reclaiming and the alkalinity or acidity of the surrounding medium exerts considerable influence on the rate and quantity of molecular breakdown during the reclaiming process. Olefin hydroperoxides can be formed during the comminution of the vulcanized scrap or during the actual reclaiming of it. The formation of active radicals, which is now considered the first step in this reaction, can occur if energy is supplied thermally. Some oxygen attack can also occur additively at the double bond to initiate the reaction, but to an insignificant degree only, and therefore it need not be reflected in any change in unsaturation of the reclaim. There is no doubt that reclaiming conditions permit these reactions to take place. The data have shown that the plasticization of the vulcanized scrap cannot be attributed to outright oxidation of the reclaim. Furthermore, peroxidic reactions, if once initiated, can continue even when reclaiming has been interrupted after a short time interval. Indeed, short-interval reclaiming should produce more unstable reclaims on the basis of this reasoning. Also, a hydroperoxidative breakdown of the vulcanized rubber during reclaiming should produce a reclaim of high and unchanged unsaturation, unless the time interval of the reaction is carried to extremes, in which case secondary reactions take place. The initial high rate of molecular breakdown, as seen in pan-process reclaiming, can be explained on the basis of peroxidative chain reactions where, after the comparatively short time interval necessary for the initial formation of radicals, a high rate of activity can be expected. These peroxides are thermally unstable, and as the time of reclaiming increases a great number of active radicals are formed which, in turn, may react with each other, causing mutual inactivation.

The great differences observed between the alkali- and metallic chloride-digested reclaims can be explained on the same basis. In acid media the decomposition of olefin peroxides was found to proceed largely along the lines of triol formation, which would not call for any loss in unsaturation. (The similarity between the data obtained from reclaims digested in water and those digested in metallic chloride can be explained by the fact that both are obtained from acid media.) In an alkaline medium, however, the decomposition of the hydroperoxides proceeds much farther and at a

greater rate than in an acid medium. This is reflected in the high rate of molecular breakdown during the initial reclaiming periods of the alkaline reclaim. Furthermore, reactions between olefins and olefin peroxides are favored thermally and could result in a decrease of unsaturation, as occurs during the longer reclaiming periods in alkaline medium. This reaction would not result in additional molecular breakdown, which explains the chloroform-extract constancy in the later period of reclaiming intervals.

The Structure of the Polymer and Its Influence on Reclaiming

The considerations and data mentioned above refer to the natural-rubber polymer and its reactions during reclaiming, without the addition of oils and catalysts. It was soon found, however, that the synthetic polymers, including butadiene-styrene and butadiene-acrylonitrile copolymers as well as polychloroprenes, were not as responsive to reclaiming as natural rubber. Figures 11 and 12 show the change in plasticity obtained in general from vulcanized GR-S and natural rubber respectively. While the initial breakdown with GR-S is very pronounced, resulting in a high plasticity over a short time interval, longer reclaiming intervals appear to reharden the polymer. In contrast thereto, vulcanized natural rubber scrap breaks down very fast initially and remains more or less constant in its plasticity increase. Some hardening can occur; however, this is so small that it is often unobservable.

The heat-hardening of GR-S polymers has been well known to the compounder for some time. GR-S compounds which were subjected to artificial ageing tests demonstrated this phenomenon; but no initial softening of these compounds was noticed at the time. This is not surprising, because artificial ageing tests are usually carried out in an oxygen-rich atmosphere for more than half an hour, whereas reclaiming is carried out in an oxygen-deficient atmosphere. This difference in the behavior of GR-S and natural rubber will cause the difficulties encountered in the reclaiming of mixtures of these polymers or of mixtures of natural rubber with any of the other two synthetic polymers mentioned. Since the above experiments were carried out in the absence of any reclaiming oil and catalysts, the results must be attributed to polymer structures and behavior under reclaiming conditions. While the structure of natural rubber had been fairly well known over a period of years, that of GR-S had to be ascertained before any conclusions could be drawn as to its influence on the reclaiming process.

The ultimate general behavior of any polymer depends on a number of features; the chemical make-up of the monomers from which it was prepared and the physical state and structure of the polymer chains are
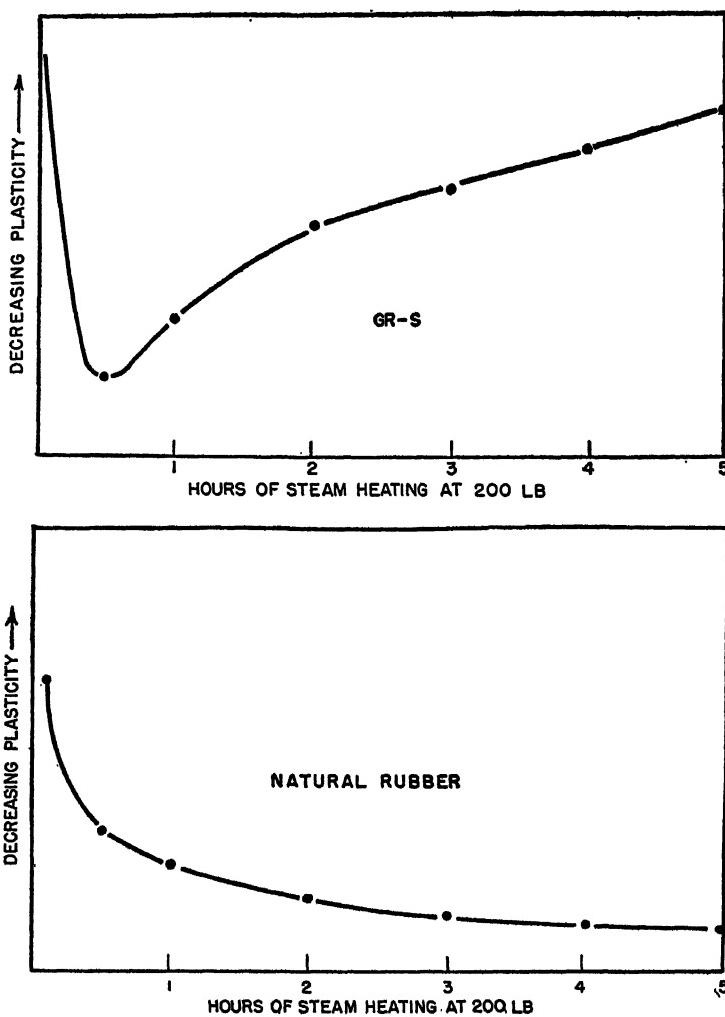
dipole linkage between the molecular chains. These groups may be intro-
duced, at least temporarily, during the reclaiming reaction.

X-ray analysis reveals that natural rubber polymer contains the iso-
prene units in the cis-position of the 1, 4-addition. The distribution of
the molecular sizes is wide, the chains are coiled tightly and yet free
enough to permit slippage with comparative ease upon application of
stress. The following formula is a schematic representation of the polymer
chains:

$$-CH_2-\overset{\overset{\displaystyle CH_3}{|}}{C}=CH-CH_2-\left|-CH_2-\overset{\overset{\displaystyle |}{C}}{\underset{\underset{\displaystyle CH_3}{|}}{}}=CH-CH_2-\right|-CH_2-\overset{\overset{\displaystyle CH_3}{|}}{C}=CH-CH_2-$$

Isoprene unit
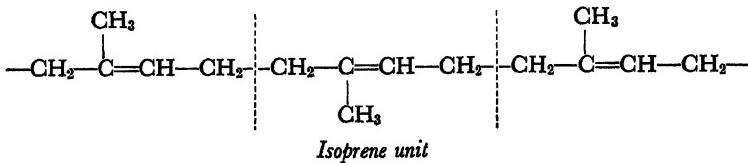
It can be expected that chain breakdown would occur at those points
where the smallest amount of energy has to be expended to accomplish it.
The breaking of a normal C—C bond requires 81 kilocalories, whereas
a double bond (C=C) requires 145 kilocalories. However, in calculating
such energy requirements for a particular system, the resonance of the
fragments formed in the breakdown of the polymer molecule must be
taken into consideration. Calculations have shown that resonating sys-
tems have a greater total energy than that which could be accounted for
by considering only the bond energies of the normal structure. This
additional energy amounts to 19 kilocalories for the allyl radical $CH_2=$
$CH-CH_2$). These energies cannot be neglected when the heat of fission
is calculated. To break the bond at the α-carbon atom in the rubber
chain,

$$-CH_2-\overset{\overset{\displaystyle |}{C}}{\underset{\underset{\displaystyle CH_3}{|}}{}}=CH-CH_2-\left|-CH_2-\overset{\overset{\displaystyle |}{C}}{\underset{\underset{\displaystyle CH_3}{|}}{}}=CH-CH_2-\right.$$
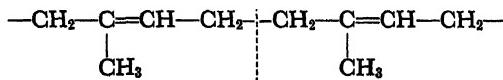
only 43 kilocalories would be required, since two allyl radicals would be
formed (81 − 38 = 43).

The unsaturation of the reclaims prepared from natural rubber is
hardly different from that of the original vulcanized scrap. This can be
expected if the molecular breakdown occurs along the lines indicated
above and the process would have to be classified as depolymerization.
Naturally, smaller chain segments can be expected to show greater plas-
ticity than the original scrap. As the reaction proceeds over a period of
time, it is possible that chain fragments reacting with each other can
cause a decrease in plasticity. It should be mentioned here that while
the above picture is that of a crude, nonvulcanized polymer, vulcanized
scrap will be somewhat different in structure and breakdown. Vulcaniza-
tion introduces some primary valency bonds, apart from causing other

changes of a physical and structural character. Thus, we can expect to find some rigidity imparted to the system, which did not exist in the crude polymer. However, few primary valency bonds are introduced in this way, and therefore the overall picture of the structure of the polymer chains will not be changed to any great extent. Furthermore, reclaiming does not remove these sulfur linkages.

The variety of changes introduced into the polymer chains during polymerization of a synthetic polymer is by far greater than those introduced by sulfur linkage. If, for example, isoprene is polymerized synthetically, 1,2-addition as well as 3,4-addition, as schematically represented, below, can occur, besides the 1,4-addition.

$$-CH_2-\underset{\underset{\parallel}{\overset{|}{CH_2}}}{\overset{\overset{|}{CH_3}}{\underset{|}{C}}}-CH_2-\underset{\underset{\parallel}{\overset{|}{CH_2}}}{\overset{\overset{|}{CH_3}}{\underset{|}{C}}}-CH_2-\underset{\underset{\parallel}{\overset{|}{CH_2}}}{\overset{\overset{|}{CH_3}}{\underset{|}{C}}}-$$

Isoprene unit

$$-CH_2-\underset{\underset{\parallel}{\overset{|}{CH_2}}}{\overset{|}{\underset{|}{CH}}}{\overset{}{\underset{C-CH_3}{}}}-CH_2-\underset{\underset{\parallel}{\overset{|}{CH_2}}}{\overset{|}{\underset{|}{CH}}}{\overset{}{\underset{C-CH_3}{}}}-CH_2-\underset{\underset{\parallel}{\overset{|}{CH_2}}}{\overset{|}{\underset{|}{CH}}}{\overset{}{\underset{C-CH_3}{}}}-$$

Isoprene unit
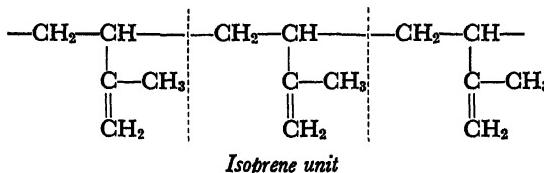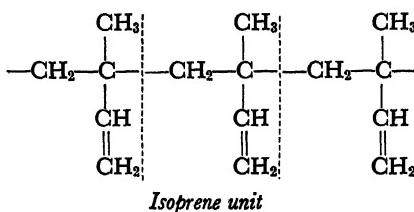
Both result in branched chains; however, the substitutional methyl group has a different position in these chains, which would influence depolymerization reactions. If butadiene is polymerized, only two kinds of addition products are obtained because this monomer does not possess any substitutional methyl group. The following is a schematic presentation of the 1,4- and 1,2-addition products respectively:

$$-CH_2-CH=CH-CH_2-|-CH_2-CH=CH-CH_2-|-CH_2-CH=CH-CH_2-$$

Butadiene unit

$$-CH_2-\underset{\underset{\parallel}{\overset{|}{CH_2}}}{\overset{|}{\underset{|}{CH}}}-|-CH_2-\underset{\underset{\parallel}{\overset{|}{CH_2}}}{\overset{|}{\underset{|}{CH}}}-|-CH_2-\underset{\underset{\parallel}{\overset{|}{CH_2}}}{\overset{|}{\underset{|}{CH}}}-$$
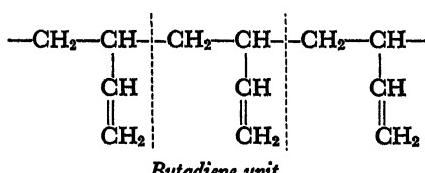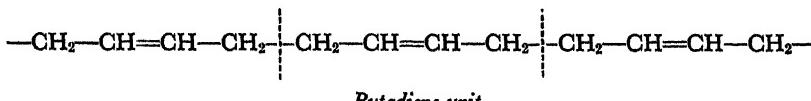
Butadiene unit

The 1,4-addition product results in long chains as found in natural rubber; the 1,2-addition product forms branched chains. Infrared spec-

trography has shown that in the GR-S polymer about 19 per cent of the butadiene units are linked in the 1,2-position.

Ozone analysis has indicated that about 31 per cent of the styrene present in GR-S forms a copolymer, where butadiene and styrene alternate and where the butadiene is present in the 1,4-position. Of the styrene present in GR-S, 40 per cent forms molecules where two styrene units are paired by polymerization. These pairs are separated by one or more butadiene units in 1,4-position. The remainder of the styrene enters into branched molecules formed by 1,2- and 1,4-addition polymerization of butadiene units. Some evidence has been presented that styrene may also polymerize by forming a methyl group outside of the chain. While natural rubber presents a rather simple, repetitive and orderly fashion of polymerization, all of the above-mentioned structural chain units occur in GR-S. In natural rubber the groups which are of importance in the reclaiming reaction are spaced at regular intervals and can be considered identical with each other in regard to the possibility of the reactions occurring throughout the molecular chains; synthetic rubber with its conglomerate of chain structures will permit a variety of reactions to occur simultaneously during the reclaiming process, and will make their control much more difficult even under identical and controlled thermodynamic conditions.

Linear polymers having thermal double bonds in the side chain will react differently under reclaiming conditions than polymers containing the double bonds in the main chain. The absence of double bonds in the main chain will not permit the peroxidative breakdown as encountered in natural rubber, and therefore the reclaiming of such polymers will be adversely affected. At the same time, however, the double bonds in the side chains will be capable of reaction (the formation of carbonyl, epoxide and hydroxyl groups is entirely possible) and recombination with other side chains or with chain fragments obtained from the 1,4-addition polymerization will result in network-like structures. The result will be a hardening, rather than a softening, of the polymer during progressive reclamation.

Studies on the breakdown of GR-S compounds at high temperature in oxygen-containing atmospheres have shown that a preliminary breakdown occurs before hardening of the polymer sets in. These effects have been ascribed to a scission of chains occurring predominantly in the early stages of heating, which is then overshadowed by fusion of the chains predominating in the late heating stages. Going back to the time-plasticity curve for GR-S (Figure 11), it can be deduced that similar breakdown effects exist during the reclaiming of GR-S. The original, pronounced but temporary increase in the plasticity can be attributed to scission of the polymer chains. Later on, as more and more active

radicals form, secondary reactions will set in and cause the recombination of fragments and side chains, resulting in a less plastic network structure. Therefore, the heat-hardening will be more pronounced in the later stages of reclaiming. Thus, the two diametrically opposed directions of the breakdown reactions can be held responsible for the peculiar shape of the GR-S plasticity curve. As the temperature of reclaiming is decreased, the shape of the curve will change to a certain extent; the increase in plasticity during short reclaiming time will not be as marked, but neither will heat-hardening progress at such a rapid pace.

This behavior of GR-S has caused considerable trouble wherever fiber removal is carried out by fiber hydrolysis during the reclaiming of the polymer. Considerations of an economic nature, as well as those concerning the corrosion of the pressure vessels, unfortunately result in a combination of circumstances which would make a longer reclaiming cycle advantageous for production. However, as the heat-hardening of GR-S progresses, the difference between the plasticities of natural rubber and GR-S in a mixture of the two becomes more and more pronounced. When milling and refining of this inhomogenous mass is attempted, after drying of the digested stock, proper mixing of the two components cannot be achieved and the hard particles fall off the refiner as tailings. This condition is still more aggravated in the presence of caustic, because an alkaline medium increases the rate of heat-hardening of the GR-S considerably.

The other two synthetic polymers available to any great extent for reclaiming, namely the butadiene-acrylonitrile copolymer and polychloroprene, react very similarly in regard to heat-hardening. Apparently, however, plasticization of these polymers is more of a problem, and neoprene, in particular, is very difficult to handle. Determination of unsaturation becomes rather inaccurate under these conditions and therefore very little information is available about the reactions occurring during reclaiming.

The Effect of Reclaiming Agents and Catalysts

Very little is also known as yet about the effect of reclaiming agents and catalysts and the reactions which occur in their presence.* Originally it was assumed that any plasticizer for the crude polymer would also be a reclaiming agent for the vulcanized material. This assumption can be made only if no other function than a lubricating effect is assigned to the reclaiming oil or catalyst. Such lubrication would be effective by pushing the polymer molecules apart and changing the cohesive forces between them. It would make reclaiming oils equivalent to plasticizers

* Innumerable patents have been issued on this subject.

for the unvulcanized polymer, and would in effect postulate that any plasticizer used in the compounding of crude polymers should be a suitable reclaiming oil for the vulcanized polymers. This assumption is not true. As far as it has been now possible to ascertain, reclaiming oils and catalysts are dependent on the presence of certain chemical groups and their preferred positions within their molecular structures. Plasticizers effective for the crude polymer are not necessarily reclaiming agents for the vulcanized form. This holds particularly true for the synthetic polymers.

Later on the assertion was made that mixtures of natural and synthetic rubber could be reclaimed successfully, provided that both were swollen to the same extent by a given plasticizer. Again this proved fallacious. No connection has been found between the heat of swelling of rubber in various solvents and the effectiveness of the latter as reclaiming oils, nor have practical reclaiming procedures confirmed the above assertion in any way. In general it can be said that:

(1) Reclaiming agents and catalysts which are effective in the reclaiming of the synthetic polymers will also be effective for natural rubber. However, not all agents effective for natural rubber are useful for synthetic types.

(2) It is not essential for a reclaiming agent to remain in the polymer after the reclaiming process.

(3) Certain chemical groups active in the reclaiming reaction have been recognized, but none of them has as yet been fully evaluated. Some of the reclaiming agents are capable of preventing or delaying the heat-hardening reaction of the synthetic polymers. Others merely increase the rate and extent of molecular breakdown, but permit heat-hardening to occur.

One of the reclaiming agents for natural rubber which has been widely used over a period of years is coal-tar solvent naphtha. It does not remain in the polymer after the reclaiming process has been carried out. It exerts considerable influence on the amount and rate of molecular breakdown (Figure 13), but at the same time has very little effect on the acetone-extractable matter of the polymer. Molecular breakdown is again enhanced in heater-reclaiming as compared with that occurring in digester-reclaiming. The presence of solvent naphtha results in decrease in unsaturation, which is greater if smaller quantities of solvent naphtha have been used (Figure 14). This fact has been explained on the basis of the swelling caused by the solvent naphtha, resulting in progressively greater polymer chain separation as the quantity of solvent is increased. Keeping in mind the important role oxygen apparently plays during the reclaiming process, it can be expected that small quantities of solvent

D. S. LE BEAU

From le Beau, India Rubber World (October, 1948)
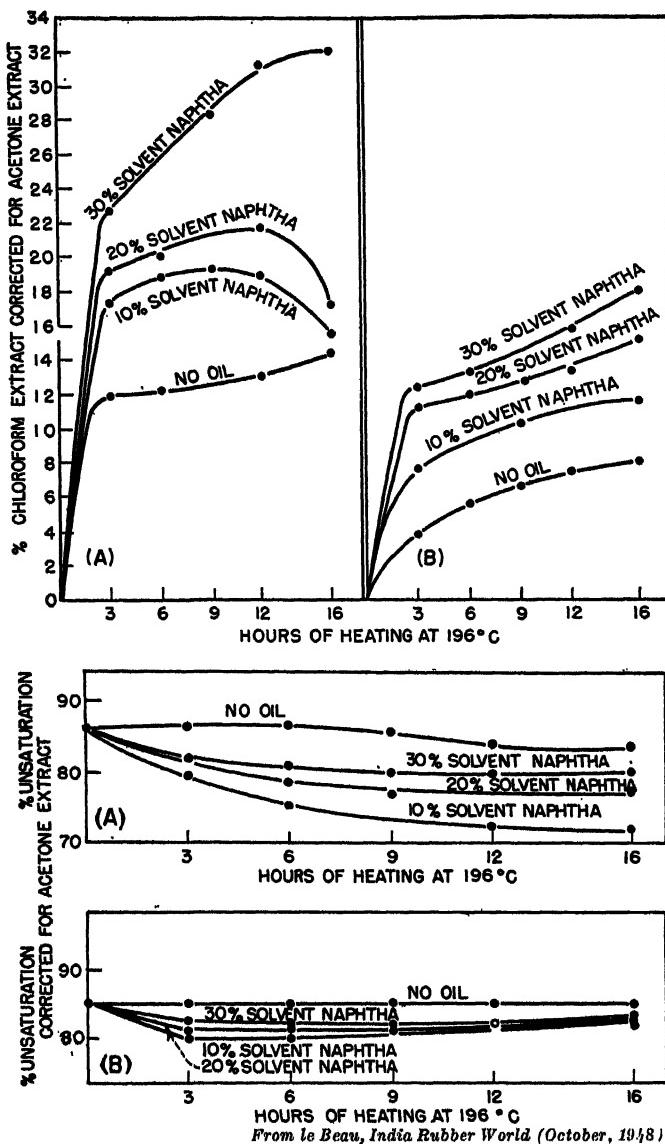
Figures 13 and 14. The effect of solvent naphtha on (A) heater reclaiming and (B) digester (water) reclaiming.

naphtha will increase the access to oxygen by swelling the polymer and thereby prying the molecular chains apart. However, the chains will not be pried apart to any great extent if the quantity of solvent naphtha is small, nor will there be enough of it to block the space between the chains uniformly. This nonuniformity will cause some strain and defor-

mation on the part of the rubber molecules. Oxidation will be favored under the above condition and these effects can be expected to be more pronounced in the heater than in the digester process. As increasing portions of solvent naphtha are used, the chains are pried farther apart and the distribution of the solvent naphtha molecules in the rubber becomes more uniform. Therefore, reactions between the chains are less likely to occur because chemical forces become weaker as the distance between the rubber molecules increases. Also the solvent naphtha molecules can be expected to exert a blocking effect. A chemical combination
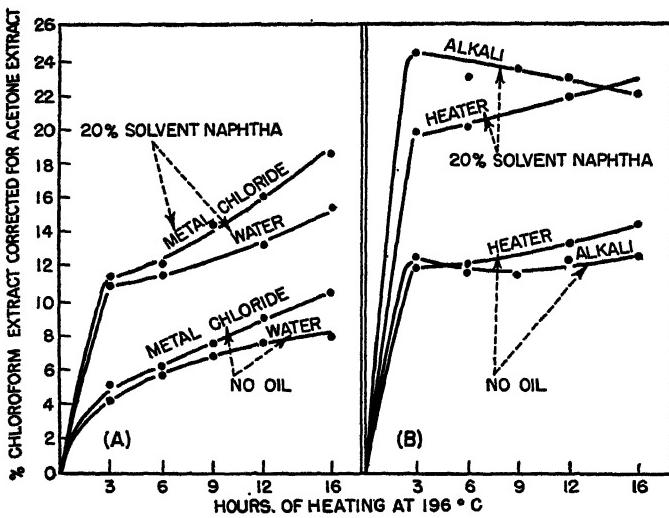
From le Beau, India Rubber World (October, 1948)

Figure 15. The influence of solvent naphtha and defibering agents on reclaiming.

between the rubber and solvent naphtha probably does not take place, because the acetone extract of the natural rubber reclaimed in the presence of various amounts of solvent naphtha is almost constant; if such a combination had occurred, the decrease in unsaturation would have been greatest when more solvent naphtha was used.

It is interesting to compare the results obtained with and without solvent naphtha when reclaiming is carried out either in an acid or in an alkaline medium (Figure 15, 16). The effect of the surrounding medium (heater or digester) and its acidity and alkalinity still predominate in guiding reclaiming reactions. Although the presence of solvent naphtha enhances the reclaiming reactions, it apparently does not change their course; hence, it can be considered a catalyst for these reactions. No such experimental data are available for the reclaiming of GR-S with solvent naphtha. However, considerable solvent naphtha has

to be used in the reclaiming of GR-S, and the resultant product has such poor properties that it is doubtful whether solvent naphtha can be considered a useful agent for GR-S.

A study of the effect of alkylated phenolsulfides * on the reclaiming of GR-S indicated that these reagents are catalysts for the oxidative scission reactions. The chemical structure of these compounds was found to have considerable influence on their reclaiming capacity. Isomeric differences in the structure accounted for minor differences in reclaiming properties; however, substitution on the phenol ring had a strong bear-
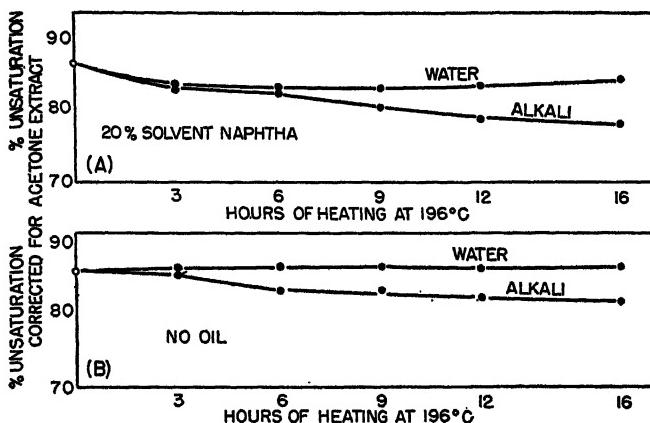
From le Beau, India Rubber World (October, 1948)

Figure 16. The effect of solvent naphtha on reclaiming.

ing on the reactivity of the compounds. Xylenol sulfides were found more active than cresol sulfides and a nonsymmetrical trimethyl phenol sulfide was found to be most active of all. If a para-substituted alkyl group was present, its chain length did not appear to affect its reclaiming properties. If, however, two or more alkyl groups were introduced into the phenol nucleus, the reclaiming activity of these compounds was enhanced. The position of the alkyl groups was not as significant as their number. In contrast thereto, the position of the sulfide groups on the phenol ring was found to be very important; greater reclaiming reactivity resulted if these groups were placed either in the ortho or in the para position. It was also found that compounds having substitution of alkyl groups on the phenol ring showed better reclaiming properties than those containing fused-ring phenolic structures. Phenols which did not contain any substitutional alkyl groups had decreased reclaiming activity, and the introduction of additional hydroxyl groups in the phenol ring also decreased this activity. It was at first thought possible that the effect

* See U. S. Patents 2,193,624 (1940); 2,333,810 (1943); 2,359,122 (1944); 2,363,873 (1944); 2,372,584 (1945); 2,469,529 (1949).

of alkylated phenol sulfides might have been brought about by a monosulfide-polysulfide equilibrium. Experimental data have, however, shown that this mechanism is not the cause for their reclaiming activities. It is known that alkylated phenols protect synthetic rubbers during the ageing process, and it is possible that their presence during the reclaiming process may retard the recombination of chain fragments in a similar way.

Aliphatic and aromatic mercaptans * have, at times, been considered useful reclaiming agents for GR-S. The aliphatic mercaptans, however, are rather poor reclaiming agents, regardless of the length of the aliphatic chain. The aromatic mercaptans were found to be more effective, particularly if the phenyl group was replaced by the naphthalene group.

The third group of chemicals which has been deemed generally and extraordinarily effective in reclaiming are amines. Whereas some of the other reactive reclaiming agents are not applicable to neoprene or butadiene-acrylonitrile polymers, the amines † are effective for the reclaiming of these polymers. Primary amines gave better results than secondary; and the latter were more effective than tertiary amines. In contrast, it was found that quaternary ammonium bases or their hydrochloric salts were not able to compete in effectiveness. Aliphatic amines were more useful in reclaiming than aromatic amines. The introduction of hydroxyl, nitro, or sulfonic groups decreased reclaiming ability. The influence of these substitutional groups is similar to that mentioned when discussing the structure of substituted alkyl phenol sulfides.

As can be seen from the preceding paragraphs, our knowledge of the effect of various chemical groups on reclaiming is sparse. Some of the above groups of reclaiming oils can enhance peroxidative cleavage of the polymers; however, it would be premature to designate this as the only reaction occurring during the reclaiming process. Amines, for example, have been found to prevent or delay the heat-hardening of the synthetic polymers, an effect which would have to occur quite apart from the peroxidative scission. It is interesting to note that amines which can be used as reclaiming agents often are also useful antioxidants for the polymers.

Structure, Properties and Evaluation of Reclaim

While the amount of chloroform extract may give a good theoretical basis for the evaluation of the amount of breakdown which has occurred during the reclaiming process, it does not explain the mode of molecular breakdown. Undoubtedly, the type of molecular breakdown will affect

* See U. S. Patents 2,211,592 (1940); 2,414,145 (1947); 2,467,789 (1949); 2,471,866 (1949).
† See U. S. Patents 2,304,548-51 (1942); 2,338,427 (1944); 2,361,015 (1944); 2,414,428 (1947); 2,423,032 (1947); 2,477,809 (1949); 2,478,826 (1949).

the properties of the reclaim. Theoretically, the most desirable break-down would be one where the polymer chains are only shortened, with-out being subjected to netting. In this case the chloroform extract would be a true measure of the molecular breakdown and its viscosity would give some indication of the molecular weight of the soluble fraction. However, it should not be forgotten that in the case of a reclaim polymer certain changes in the shapes of the molecules have been introduced by vulcanization, resulting in the partial formation of a network by virtue of primary valency forces. Therefore, the molecular chains contained in the chloroform extract may be branched to some extent and any molecular-weight determination on the basis of viscosity measurements will not be quite true. If, at the same time, netting of the polymer should also occur during the reclaiming process, the chloroform extract may contain polymer chains of highly branched or netted structure and of corpuscular shape. Such particles would, in general, be quite insoluble, and therefore it can be expected that either the molecular chains which are present in the chloroform extract do not contain a great deal of netting and branching, or that the particles are of a low enough molec-ular weight to remain soluble.

Similar considerations hold true for the chloroform-insoluble part of the reclaim. If netting does not occur during the reclaiming process and the original polymer molecules were predominantly long chains without branching or netting, the chain fragments present in the reclaim will be shorter but of the same general shape as the original chains. If netting occurs during reclaiming or if the original polymer molecules were highly branched and netted, the chain fragments present in the reclaim will be of very corpuscular shape.

While it might appear, offhand, that such considerations are of a theo-retical nature only, it must be recognized that the kind of breakdown occurring during the reclaiming process and the recombination of chain fragments will determine the useful properties of the reclaim. It has been pointed out in the beginning that reclaim is used even when little or no outright economic advantage is involved, because of the helpful process-ing properties which it can impart to the various rubber compounds. These properties, however, will be a function of the reclaiming con-ditions. For example, it is possible to produce a reclaim of short and rather unbranched chain fragments, capable of considerable recovery. This reclaim possesses high shrinkage and plasticity and has its specific usefulness. In contrast, a reclaim possessing corpuscular chain fragments will show high dimensional stability, yet it can also possess high plasticity if the corpuscular chain fragments are small enough or if adequate amounts of permanent plasticizer are present. Furthermore, the physical

properties of both reclaims, such as tensile strength and elongation, can be identical.

The measurement of plasticity has of late been increasingly used as a means of evaluation and control of the quality of the reclaim. The above considerations indicate this to be erroneous. Furthermore, a highly non-uniform reclaim may show identical plasticity values with a very uniform reclaim, provided that "enough mortar is available to carry the bricks," or in other words, a very wide and uneven chain-fragment size distribution need not necessarily affect the plasticity measurement. While literature shows that plasticity measurements have been taken as an indication of the effectiveness of a reclaiming process, it is evident that such values will often not indicate whether or not netting has occurred during reclaiming, or whether a given reclaim will be capable of forming and maintaining a self-supporting film under stress. Possibly a coherence factor will have to be introduced into reclaim evaluation.

The physical properties of vulcanized elastomer polymers are lowered during the reclaiming process. The rate of this decrease is most pronounced during the very first minutes of the reclaiming treatment and levels off fairly soon. This is not surprising, as the data have indicated that the rate of molecular breakdown is greatest during the initial period of reclaiming. The physical properties will naturally be most affected whenever the decrease in average molecular size is greatest. This condition will be fulfilled in the original stages of reclaiming. Later on we can also expect that chain fragments may be mutually inactivated by recombination. Attempts to improve the tensile strength and abrasion properties by copolymerization of the reclaim with appropriate monomeric compounds have been successful.

The results of the reclaiming process will naturally depend on the temperature at which it is carried out. Interestingly enough, the presence of reclaiming oils is actually beneficial in preventing some of the degradation of physical properties. This is particularly true for the elongation. Inasmuch as we have no means available to measure the size and shape of the chain fragments of the chloroform-insoluble part of the reclaim, we have no means to determine whether this action of the reclaiming oil is due to a change in the size of the chain fragments or to less netting. X-ray analysis, which might provide an answer, has so far not been very successful; however, insufficient work has been done along these lines. Natural rubber reclaim in the unvulcanized and stretched state will not show a fiber pattern even though its elongation at break is beyond that at which fiber structure can be observed in vulcanized rubber. Continuous flow during exposure to x-rays may possibly explain why the rubber molecules do not align properly and interlock at certain positions. This behavior can be expected from a low molecular-weight polymer. How-

ever, if a compound is prepared from natural rubber reclaim and vulcanized very slightly, elongations up to 750 per cent can be obtained. X-ray diffraction analysis of this compound yields indications of a fiber pattern. It is known that upon air-oven ageing of vulcanized rubber, greater elongations are necessary to obtain the fiber pattern than for an identical unaged vulcanized compound. This effect can possibly be explained by molecular breakdown and resulting flow or by severe oxidation which hampers the fiber alignment. Infrared analysis has shown that reclaim does not consist of severely oxidized rubber; however, the alignment of the molecules could also be hampered by branching, netting, or cyclization. At this time, it is impossible to decide which of the above possibilities may cause the phenomenon.

REFERENCES

1. Andrews, and Tobolsky, A., *J. Applied Phys.*, 17, 352 (1946).
2. Ball, J. M., "Reclaimed Rubber," New York, Rubber Reclaimers Assoc., Inc., 1947.
3. Bateman, L., and Koch, H. P., *J. Chem. Soc.*, 600 (1944).
4. Boland, *India Rubber J.*, 108, 405 (1945).
5. ———, and Orr, G. H., *Trans. Inst. Rubber Ind.*, 21, 133 (1945).
6. Busse, W. F., *Ind. Eng. Chem.*, 24, 140 (1932).
7. Carpenter, A. S., *Ind. Eng. Chem.*, 39, 187 (1947).
8. Cook, W. S., *et al.*, *Ind. Eng. Chem.*, 40, 1194 (1948).
9. Cotton, F. H., *Trans. Inst. Rubber Ind.*, 6, 487 (1931); 21, 153 (1945).
10. Davis, C. C., and Blake, J. T., "The Chemistry and Technology of Rubber," New York, Reinhold Publishing Corp., 1937.
11. Dogadkin, B., *Rubber Chem. Tech.*, 22, 37 (1949).
12. Farmer, E. H., *Rubber Chem. Tech.*, 16, 17, 45 (1943).
13. ———, *India Rubber J.*, 112, 119 (1947).
14. ———, and Michael, S. E., *J. Chem. Soc.*, 513 (1942).
15. Flory, P. J., *J. Am. Chem. Soc.*, 69, 2893 (1947).
16. Fuller, C. S., *et al.*, *J. Am. Chem. Soc.*, 62, 1905 (1940).
17. Kilbourne, F. L., Jr., and Miller, G. N., *Ind. Eng. Chem.*, 22, 69 (1930).
18. Kirby, W. G., and Steinle, L. E., U. S. Patent 2,279,047 (1942).
19. le Beau, D. S., *Rubber Age*, 62, 51 (1947).
20. ———, *India Rubber World*, 118, 59 (1948); 119, 69 (1948).
21. Palmer, H. F., and Kilbourne, F. L., Jr., *Ind. Eng. Chem.*, 32, 512 (1940).
22. Shepard, N. A., *et al.*, *Ind. Eng. Chem.*, 20, 143 (1928).
23. Sverdrup, *et al.*, Paper presented at the Rubber Div. Am. Chem. Soc. Meeting in Detroit, Nov., 1948.
24. White, L. M., *et al.*, *Ind. Eng. Chem.*, 37, 770 (1945).
25. ———, *Rubber Chem. Tech.*, 18, 826 (1945).
26. Winkelmann, H. A., *Ind. Eng. Chem.*, 18, 1163 (1946).
27. Winn, H., *et al.*, *Ind. Eng. Chem.*, 38, 71, 1052 (1946).

CHEMICAL AND PHYSICAL MODIFICATIONS OF TEXTILE FIBERS

H. C. Borghetty

General Dyestuff Corporation, New York, N. Y.

EVER SINCE the successful introduction of synthetic fibers, there has been a constant increase in activity towards the development of chemically-modified cellulosic fibers. This in turn has given the necessary impetus, long overdue in the textile industry, to scientific workers of various other fields and has permitted the pooling of information that has brought about the many investigations and practical applications now under way.

Although the textile industry actually pioneered the great industrial revolution of the eighteenth century and continued to be alert to all new mechanical tools, it did not show the same inclination to participate in the improvements made available by the chemical and physical discoveries of the twentieth century. As the reasons for this lack continue to exist and will therefore continue to impede some of the potential developments, they should be recognized as follows:

(1) Textile fibers are numerous, derived from many sources, mineral, vegetable and animal and thus are subject to various methods of dyeing and finishing. This resulted in the establishment of many different branches of the industry in different geographical locations, causing considerable difficulty in cooperation and coordination.

(2) The *art* of fiber- and fabric processing is rather intricate and requires a painstaking and lengthy apprenticeship that in turn discouraged the entry of the chemist into the textile mills until a few decades ago. These chemists are now fulfilling the coordinating function with the various interested industries, which promises much.

(3) The system of marketing fabrics is so complex that the actual product is oftentimes not owned at the source where the greatest amount of workmanship is performed. Unfortunately, also, the greatest profits are reaped where less scientific knowledge is re-

quired, which means that often no funds for scientific research are made available.

The future advancement of the industry, in spite of these drawbacks, is assured mainly because of the renewed interest that is evident and which is traceable to the foundations laid by the following related and cooperating service industries:

(1) *The dyestuff industry* supplies the textile manufacturers with the colors so essential for the beautification and styling of modern garments. This branch of the chemical industry made such rapid strides in a few decades that it greatly outpaced the dyeing techniques available and then paused, awaiting wider application of the many dyestuffs developed. Dye application techniques, meanwhile, are being so improved by the use of new continuous textile machinery that the dye manufacturers are again readjusting their programs.

(2) *The plastics industry* originally supplied coating lacquers for fabrics used for raincoat cloths, artificial leather and bookbindings. The multitude of new plasters now supplement and supplant textile fibers in many uses. Plastics such as Saran, Vinyon and nylon are full-fledged textile fibers and bridge the gap between the textile and plastic fields, thus yielding efficient cooperation. This new industry is thoroughly scientific and has been indispensable in the new textile recovery.

(3) *The synthetic fiber industry* is completely new, fully integrated, and controls not only the raw material supply, but also the entire chemical and manufacturing process from beginning to end, with a definite interest in marketing its own product. This industry was admirably suited to direct its spectacular growth along sound chemical engineering principles.

Man-made fibers possess the advantage over natural fibers of being produced in a chemical state almost completely free of impurities, whereas the natural fibers contain many impurities, dirt, grease, etc., that must be removed before spinning and weaving. The goal of the synthetic fiber manufacturer, however, is to produce materials possessing the qualities of natural fibers, hence the desire to make a rayon resembling silk. At one time rayon was called *artificial silk*, indicating the hope that the goal had been achieved. The similar terms, *artificial wool* and *protein fiber*, were coined to indicate substitution of the animal fibers.

The synthetic fiber industry now supplies approximately 17 per cent of the total textile output. Its close contact with every phase of research, manufacturing and merchandising has stimulated the entire textile field.

Types of Modifications

Chemical modifications had the primary object of varying the dyeing affinity of cellulose fibers in an endeavor to render them receptive to wool dyestuffs. Such a goal would mean greater flexibility of coloring; would permit simultaneously dyeing of wool and of chemically-modified cotton in one bath, thus enabling a "covering up" of the less expensive cotton component in mixed fabrics. A means to achieve this is indicated [1] by the use of stearylamine acetate in conjunction with formaldehyde. The use of benzoyl chloride to render cotton immune to dyeing is an application giving a two-tone effect.

Physical modifications of fibers formed the basis of textile finishing operations and comprised mechanical methods of varying the surface characteristics of fabrics: calendering, schreinering napping, shearing, embossing, glazing, tentering, etc. were the old-time trade secrets of the practical textile finishers, and were later supplemented by the wet processing operations, such as starching which imparts firmness, "feel" and bulk to fabrics, or oiling with solutions of sulfonated olive oils which impart softening and draping qualities. These materials which together with hygroscopic agents, waxes, and dulling agents, are water-soluble and therefore removable during the first washing of the garment by the consumer, constituted the majority of the chemicals applied to cloth until the introduction of synthetic resins which impart permanent modification of physical characteristics. One exception in the former methods, which constituted a milestone in the finishing of textiles, was the permanent modification of cellulose obtained by successful adaptation of Mercer's idea of swelling the cotton fiber by means of concentrated sodium-hydroxide solution. This treatment gives increased tensile strength and increased affinity for dyestuffs, as well as increased luster when applied in conjunction with tension. The mercerized fibers retain their softness and other characteristics while acquiring greater chemical reactivity. Mercerized cotton has a rounder cross section as a result of the swelling induced during the treatment.

Chemical modifications other than for the specific purpose of changing dyeing properties, impart many variations of the physical properties as well, so it is not necessary to endeavor to classify the exact borders of the changes produced. The various modifications are, therefore, best referred to accounting to the final results, rather than the exact means of achieving them. The animalization of cellulose which falls more closely under chemical changes will be covered farther on in the paper.

Cotton (Cellulose)

The uses of this fiber are so multiple because of its outstanding resistance to wear and washing, coupled with its low cost, that it maintains dominance of the textile field by showing 70 per cent of the total poundage used. Improvements possible today make it highly probable that this ratio will be maintained and that the synthetic fibers will not make greater inroads at its expense but rather at the expense of wool and silk.

A well-known permanent finish for cotton fabrics is based on the application of cellulose xanthate solution, followed by coagulation, to yield a high-luster fabric used for table cloths. Similar effects can be obtained by treatment with a cold cuprammonium solution (copper hydroxide in ammonia) that dissolves the cotton surface partially and eliminates the fuzziness; decopperizing of the fiber to remove the blue-green color gives a lustrous surface and a firmer fabric. The color can be allowed to remain and filter cloths or mildewproofed fabrics for outdoor use result. The firm finish is a consequence of the bonding effect of the dissolved cellulose during the cuprammonium reaction on the surface of the fiber. Alkali-soluble cellulose ethers, such as ethyl cellulose, can be applied and precipitated by acid to add a lustrous flexible layer of cellulose, giving a permanent finish to cotton. A parchmentized cotton surface can be obtained by rapid treatment with concentrated sulfuric acid, followed by immediate neutralization. This treatment, known as the Heberlein process, gives firm finishes of high permanence to washing.

Modification of Cellulose. The cellulose molecule is made up of many cellobiose units. Cellulose chains are arranged in bundles or micellae oriented parallel to the axis of the fiber. A group of micellae form a fibril or fiber element.

According to Mark,[2] cellulose in cotton has a molecular weight of 1,600,000, whereas other typical polymers show the following molecular weights:

Material	Molecular Weight
Cellulose in native cotton	1,600,000
Cellulose in wood pulp	1,600,000
Cellulose in rayon	80,000–160,000
Cellulose in acetate rayon	80,000–140,000
Polyamide (nylon)	25,000

Cellulose is a comparatively stable substance and its main reactions are those of the alcoholic hydroxyl groups. Its main derivatives are, therefore, ethers and esters. Although there are three hydroxyl groups to each glucose unit, the esterification of all three is difficult and often stops when one group only has been reacted. In bundles of cellulose chains, only the surface groups are apparently reacted, while the inner groups

are not reached. The mercerization process mentioned before renders the —OH group more reactive so that a water-repellent, stearyl cellulose can be obtained by reacting mercerized cotton with [3] stearamido methyl pyridinium chloride

$$H_{35}C_{17}\!-\!\overset{\displaystyle}{\underset{\displaystyle O}{C}}\!-\!NH\!-\!\overset{\displaystyle H}{\underset{\displaystyle H}{C}}\!-\!N\cdot Cl \;+\; C_6H_{10}O_5$$

$$\longrightarrow H_{35}C_{17}\!-\!CO\!-\!NH\!-\!CH_2\!-\!O\!-\!C_6H_9O_4$$
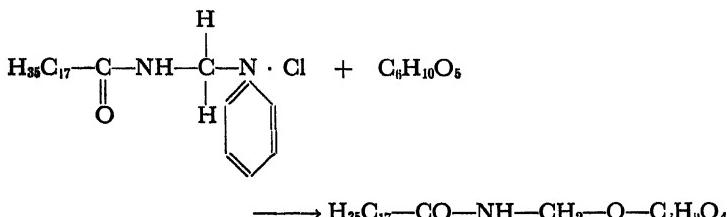
The above reaction is utilized considerably in the textile industry for cotton fabrics destined for water-repellent rain-wear garments. It points the way to a multitude of similar ones yielding permanently-modified cellulose possessing many attributes such as mildewproofing, bacteriostasis, etc. A similar reaction, also resulting in a water-repellent cotton, is accomplished by condensing N-octadecyl-N'-ethylene urea isocyanate obtained from ethylenimine, with cellulose:

$$R\!-\!N\!-\!\overset{\displaystyle O}{\underset{\displaystyle}{C}}\!-\!N\overset{CH_2}{\underset{CH_2}{\diagup\atop\diagdown}}\Big| \;+ HO\!-\!Cell. \longrightarrow R\cdot NH\!-\!\overset{\displaystyle}{\underset{\displaystyle O}{C}}\!-\!NH\cdot CH_2CH_2\!-\!O\!-\!Cell.$$
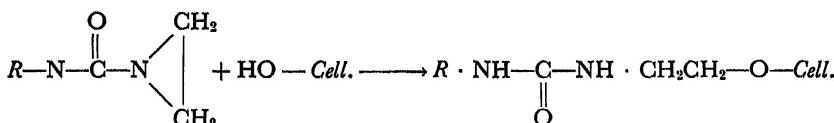
N-Octadecyl-N'-ethylene
urea isocyanate

Flameproofing has been obtained by reacting cotton with phosphoric acid and urea to yield a phosphate ester,*

$$Cell.\!-\!O\!-\!\overset{\displaystyle O}{\underset{\displaystyle}{P}}\overset{OH\cdot NH_2}{\underset{OH\cdot NH_2}{\diagup\diagdown}}C\!=\!O$$
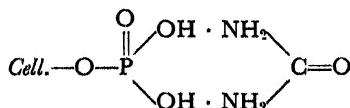
of good resistance to washing. Flameproofing, until the recent introduction of this finish, was carried out by impregnating the fibers with water-soluble ammonium phosphate, boric acid, borates and barium sulfate. Pigments such as antimony oxide, held mechanically to the fiber with binders based on chlorinated rubber, are still used extensively for tentings.

Synthetic Resins. The greatest step forward in the textile field in recent years has been the introduction of the synthetic resins or plastics. Phenol formaldehyde and urea formaldehyde have soluble monomers which readily polymerize, yielding water-insoluble polymers. Cellulose fibers are able to absorb these materials, which are then deposited permanently,

* Ban-flame, J. Bancroft and Sons Co., Wilmington, Del.

giving a variety of finishes. The wrinkle-resistant finish, both for cotton and rayon, are based upon such application, and it seems that the resin imparts substance to the fiber without in any way detracting from the feel of the final fabric. Crush-resistant properties greatly enhance cellulose, giving it some of the most desirable characteristics of the more complex wool fiber.[4]

The phenol-formaldehyde resins were later discarded for textile use because they tend to yellow and, furthermore, give a phenolic odor to the fibers; but urea formaldehyde has become one of the standard thermosetting resins, imparting crush resistance to fabrics. Melamine-formaldehyde resins have been found to be as fully satisfactory as the urea type and have a multitude of uses. Application of amino-formaldehyde resins to cotton fabrics, in conjunction with a high-temperature calendering operation to induce flow during the partial setting of the thermosetting resins, gives origin to the permanently high-luster, Ever-glaze finish used for chintzes.[5] Another application covers the use of melamine resins to give flameproof finishes on cotton.

When the resin is applied to the fabric, giving a thread locking or a bonding of the fibers instead of impregnation of the cellulose, the resulting effect is a nonslip finish,[6] which is a development of importance. Fabrics of filament rayon have always shown an excessive tendency to fray readily when very slight pressure is exerted along the warp or filling or by strain at the seams. This fraying effect is naturally caused by the very smooth surfaces of the individual threads which offer no resistance except that given by the interweave of the threads themselves. It naturally follows that the number of warp and filling threads per inch and the weave of the fabric are the two controlling factors in the case of slippage of the threads. Low-count fabrics fray so much that often they cannot be used commercially. Satin designs of 140×60 or lower exhibit great deficiency in this connection. The introduction of the nonslip finish has greatly reduced slippage, and today it is possible to manufacture additional low-count fabrics regardless of count and design, which are merchandizable.

The adaptation of the smooth, high-luster rayon satin to the lining field, in place of high-count twills, has been brought about by the application of nonslip treatments. This finish consists in bonding the warp and filling threads at their points of interweave by first saturating the fabric with a solution of a resin and then drying *while maintaining the fabric at its correct dimensions (width and length).*

The modification of this treatment gives dimensional stability to fabrics. It is well-known that cellulose fibers, because of the presence of the many —OH hydrophilic groups in the molecule, tend to swell readily in water. The resultant increase in the diameter of the yarns

gives a distortion of the threads of the fabric, causing shrinkage of the fabrics. Dimensional stability can be maintained in two ways: (1) by physical bonding of the warp and filling threads, or (2) by hydrophobizing of the hydrophilic groups to avoid swelling of the fiber. The latter effect is being increasingly developed.

Urea formaldehyde and melamine formaldehyde resins act by filling the interstices between the fibers, rendering them less apt to swell as well as giving a bonding structure. The use of formaldehyde alone blocks the majority of the surface —OH groups, rendering the fabric considerably less hydrophilic, and performs the function even more satisfactorily than the resin condensates. The introduction of glyoxal, CHO—CHO, which acts in like manner to formaldehyde but is less volatile, more reactive and requires less acid, is the basis of the Sanforset process to render rayon fabrics shrinkproof or dimensionally stable.

The above three types of resins—phenol-, urea- and melamine formaldehyde—belong to the condensation or *polycondensation* group because they are obtained by condensation, followed by passage to an insoluble polycondensate state on drying and curing, in the presence of a catalyst.

Another group of resins derived by *polymerization* or copolymerization are applied to cotton fibers to give various finishes that are rapidly replacing starch and other fugitive finishes. Such resins are the: vinyl chloride and vinyl acetate thermoplastics. They serve to give water-repellent coatings and are used as binders for pigments. Vinyl chloride-acetate copolymers are also thermoplastics. They are used [7] for the trubenizing of collar cloth in the shirting field. The plastic, applied in water emulsion, acts as a laminating agent between two pieces of fabric. Heat and pressure create bonding effect. Polyvinyl alcohol, made by hydrolysis of the acetate, is a water-soluble thermoplastic yielding clear, tough, transparent films which act as ideal sizing agents for threads and fabrics. They are also used in conjunction with glyoxal for shrinkage-resistant finishes on rayon. Styrenes made from benzene are thermoplastic types which give full-bodied finishes of good wash-fastness, while the various acrylates and methacrylates are becoming increasingly important by their ability to impart abrasion resistance, increased wearability and slip-proofing qualities when used as water dispersions. They dry to clear, nonbrittle films.

The third group of resins used widely in many phases of the textile field are the polymers represented by cellulose esters and ethers such as acetate, nitrate, and ethyl cellulose, used mainly as binding agents and to obtain modified hand of fabrics or coated effects.

New resins based on silicones have just begun to be utilized because of various properties; water repellency, slipproofing and foam killing.

These are organosilicon oxide polymers obtained from organosilicon chlorides:

$$R_2SiCl_2 \longrightarrow R_2Si(OH)_2 \longrightarrow \begin{bmatrix} R \\ | \\ Si-O- \\ | \\ R \end{bmatrix}$$

(R = alkyl)
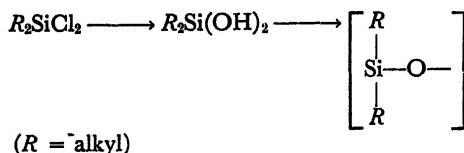
It is possible to obtain a host of varieties of finishes by the proper selection of resins and by varying the application methods. Fabrics today can be altered not only by changing the twist and weight of the yarns or the count of the fabric, but also by imparting absorbency, firmness, rigidity, resiliency, luster, coolness, and wearability. The horizons of finishing are so wide that it is easy to see how the textile chemical field is growing faster than the dyestuff field from which it originated, because while the latter imparts visual advantages, the chemical finishes embody improvement of quality, ennoblement of the properties of fibers, and increased life of the garment, and also permit the use of low-cost fibers in place of expensive ones.

Other Finishes for Cotton. Besides the standard wet-application chemicals, the resins, and the dry mechanical treatments, there are new quaternary ammonium compounds which efficiently impart a high degree of softening and also a fair amount of wash-resistance. These materials are *cationic* or reverse-charge types, a terminology used in the textile industry, to denote large colloidal molecules in which the active or *bulky ion* is the cation. Applied to the negatively-charged cellulose fiber they exhaust readily, giving a monomolecular layer of positively-charged fatty derivative on the surface. These materials, in addition, possess dye-fixing properties on many direct-dyeing (substantive) colors which heretofore, however, have also shown a simultaneous loss in light-fastness. This problem is being studied and appears to be meeting with success by the use of synthetic resins. Dicyandiamine-formaldehyde and guanidine-formaldehyde resin solutions are applied.[8] Nitrogenous bases condensed with glycerin yield similar effects.[9] Aftertreatment with polymerized amines such as polyethylene-polyamine [10] gives good dye fixation.

Synthetic Fibers

The main types of synthetic fibers are the regenerated cellulosic fibers, and to a lesser degree, the cellulose acetates. Rayon is the generic term for filaments made from various solutions of cellulose by extrusion through orifices, followed by solidification in the form of continuous filaments. The first rayon of the series was prepared by extruding an

alcoholic solution of cellulose nitrate, followed by denitrating to regenerate the cellulose—this method has now practically fallen into discard. The *cuprammonium process* has also become a minor one, yielding place to the more efficient Viscose process. Viscose rayon is produced by soaking alpha-cellulose from wood pulp in caustic soda and dissolving the aged mass in carbon disulfide. The reaction product, cellulose xanthate, a viscous liquid, is extruded through spinnerets, coagulated in dilute acid and then desulfurized to reform cellulose.

Acetate rayon, on the other hand, is the acetylated ester obtained by steeping cellulose in acetic anhydride, and then dissolving in acetone to give a spinnable solution from which the acetone evaporates after extrusion. This fiber has natural crease-resistance and lower water-absorbency on account of the diminished number of hydrophilic —OH groups and is a more highly valued textile fiber than viscose rayon. Its advantages however led to dyeing difficulties until a new series of dyestuffs were synthesized. Special dyeing procedures were required. Some methods, based on amino anthraquinones, were found to be defective in resistance to atmospheric gases (coal combination products, NO_2, SO_2, etc.) and gave the trouble known commercially as *gas fading*, necessitating the use of *gas inhibitors* or fume-proofers for correction. Alkaline solutions, amine derivatives, and guanidines have been found to remedy this objectionable feature of the blue and blue-component shades, and recent improvement of treating agents actually gives desirable protection even to washing. Formaldehyde-polyamine condensates [11] are also used.

Chemical Modifications of Cellulosic Fibers. The filamentous cellulose fibers, viscose and acetate rayon, when chopped into short lengths of 5/8 to 3 inches, yield spun-rayon or staple-rayon fibers which can be carded, spun, and woven on standard cotton-system machinery to give fabrics of entirely different aspects, feel, and warmth. This innovation started one of the foremost developments of the industry. The soft, spongy fibers that resulted were admirably suited to take the place of the finer fibers, linen and wool. Spun-rayon fabrics, coupled with the successful use of the synthetic-resin solutions and dispersions giving structural stability to the fibers, opened a new horizon of fiber development and gave the ideal physical properties to fibers.

Chemical modifications, to give fibers equal in every respect to wool, proceeded rapidly and continues to do so. Whether this is a worthwhile line of research, is not yet fully apparent but the information obtained so far appears to indicate that the knowledge acquired will be readily absorbed into the entire textile-development framework. Chemical modification of cellulosic fibers has been given a great deal of consideration in recent years and the number of patents granted on such modifications increases rapidly from year to year.

Recent modifications in cellulosic fibers actually show that many of the physical as well as chemical properties of wool can be duplicated. Physical properties such as feel, nonconductance, luster, and degree of curl render cellulosic fibers similar to wool in appearance, wearing- and warmth-giving properties. Viscose-rayon fiber at present offers the greatest possibility of success, and the effort contributed to its change is sufficiently interesting to warrant an outline of the various investigations under way and commonly referred to as the *animalization of cellulose fibers.*

Animalization of Cellulose Fibers. The primary objective of a modified cellulose having the characteristics of animal fiber, is that of rendering it receptive to wool-dyeing colors, mainly acid dyestuffs. Increased affinity for basic colors can also be imparted, but this category of coloring matters has as its only attribute great brilliancy, offset by unusually inferior light-fastness properties. Treatment of cellulose with paratoluene sulfochloride, in the presence of a solvent, yields fibers which dye readily with basic colors without need of the usual preliminary mordanting tannic-acid treatment; the affinity for direct dyestuffs is also proportionately diminished.[12]

In the more recent attempts to animalize synthetic cellulosics, appropriate consideration has been given to the practicability of achieving the results by applying the treatments at the correct stage of manufacture; hence the following two approaches: (1) Addition of animalizing agents to the mass of the rayon, prior to extrusion of the filaments; viscose rayon can be easily treated during its various stages of preparation; i.e., as alkali cellulose, xanthate cellulose, or viscose. (2) Addition of the animalizing agent after spinning of the filament: this can be achieved by direct treatment of the fiber before, during, or after coagulation. All types of cellulose-derived rayons—viscose, nitro, cuprammonium, or even acetate —can be animalized. Proteins and nonprotein, nitrogen-containing substances have been found most suitable to render cellulose wool-like, and the principal methods are described in the following sections.

Use of Proteins. The first successful animalized viscose-rayon fiber appeared in 1932 under the name of "Cisalfa." * It was produced by adding 5 per cent casein and titanium dioxide, as a dulling agent, to the alkaline solution of cellulose xanthate immediately prior to spinning, in order to lessen the contact period of the alkali. The resultant fiber was given a curliness and appeared to be quite wool-like. It could be incorporated into woolen fabrics as a cheaper component and could also be used in conjunction with regular viscose, to serve as an animal-fiber component. Several types were available in different staple lengths and deniers to blend with various grades of wool or rayon. Such fibers were superior to

* Manufactured in Italy by Cisa Viscosa.

viscose in feel, appearance and insulating qualities, but were of lower tensile strength.

Casein lends itself readily to application during viscose rayon manufacture and also forms the basis of a full-fledged casein fiber, lanital, which has a place in the economy of certain countries where sheep are scarce and milk abundant. Floxan, a pre-war, German, animalized rayon, contained up to 30 per cent casein and had strong acid-dyeing properties. Keratin, dissolved in sodium sulfide, is a satisfactory source of protein for addition to the viscose mass.[13] Fish albumin was studied in Germany by the Deutsche Eiweissgesellschaft and found to be more readily soluble than casein and more resistant to the action of the strong alkali of the viscose spinning mass. Good affinity for wool dyes results from the use of 10 to 20 per cent of this albumin. However, low contents of protein are found to be more practical, owing to the fact that the swelling of the fiber and the consequent decrease in tensile strength is in direct proportion to the protein content. The use of formaldehyde to coagulate the protein was found advantageous in controlling this swelling.[14]

Use of Nonprotein Substances, Containing Nitrogen. Stearylamine acetate has already been mentioned,[1] as a means of imparting affinity for acid dyes to cotton. Many N-containing bodies such as amines, amides and their derivatives, and quaternary ammonium salts have been used to render rayon more reactive. Primary, secondary, and tertiary amines of high molecular weight, and heterocyclic bases, e.g., pyridine, have been investigated; [15] also fatty derivatives of isatinic anhydride.

A recent method covers the use of tertiary hydroxyamines (e.g., triethanolamine) and ammonium chloride, in the presence of formaldehyde, on extruded and coagulated viscose filaments,[16] followed by curing at 300°F to transform the cellulose to:

$$Cell.\text{---}OCH_2O\text{---}(C_2H_4)N(C_2H_4O)_2\text{---}HCl$$

This approach is more in line with the reaction already indicated under the quaternary ammonium water-repellent finishing process for cotton. It creates a direct bond to cellulose and modifies the dyeing properties of the viscose, but it does not in any other manner impart wool-like qualities to the fiber.

Polyamines, obtained by reacting halogenated alkylamines with fatty amines, i.e., diethylchloroethylamine with stearylamine:

$$(C_2H_5)_2N\text{---}CH_2\text{---}CH_2\text{---}Cl + H_2N\text{---}C_{18}H_{37}$$
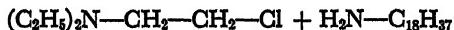
can be added to the spinning solution.[17] Reaction products of chlorinated paraffine and polyethylene polyamines have been considered. Urea is used rather often in fixing acid dyes to cellulosic fibers, particularly in printing processes, and it can also be used in viscose-stage rayon.

Cationic products based on quaternary ammonium salts are used as dye-fixing agents for acid- and direct dyes on cotton—also as auxiliary chemicals to increase the affinity of wool and fur because of the reverse charge attraction conferred by the cationics. Rayon fibers are negatively charged and the addition of a quaternary ammonium salt imparts a positive charge which attracts the negative acid-dyeing group:

$$\text{Dye So}_3^- \text{Na}^+$$
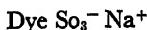
Viscose can be treated with the reaction product of trimethylamino-chloroacetic acid on stearylamine or with materials obtained by the action of betaine chlorohydrate on methylstearylamine. The resultant cellulose, however, is valued more for its hydrophobic characteristics and subsequent greater shrinkage stability than for its acid-dyeing qualities. Quaternary ammonium compounds can be used to treat rayon fibers, after coagulation, to confer reversed charge dye attraction. Materials derived by reacting benzidine dichloracetamide with long-chain aliphatic amines to give complex quaternary ammonium derivatives have been suggested [18] for this purpose. Another quaternary type derived by action of the chlorohydrin of a polyglycerin on octadecyldimethyl aniline has been used to treat rayon prior to dyeing with metachrome dyes.

Ethylenimine (and its derivatives) have been studied considerably by the I. G. Farbenindustrie A.-G. and, regardless of its difficulty in handling, appears to be destined to open new fields of investigation because of its high degree of reactivity towards —H compounds and its property of polymerizing on fibers. Added directly to xanthate cellulose, it forms complex nitrogen and sulfur derivatives of cellulose or mercaptothioazolidin ethers. Ethylenimine reacts in the gaseous state during the extrusion of the viscose filaments, prior to coagulation, giving surface effects. Ethylene ureas which still possess the imine ring and its accompanying reactivity can be formed by reacting ethylenimine with isocyanates. These products can be used with hydroxylic compounds, such as cellulose, to give:
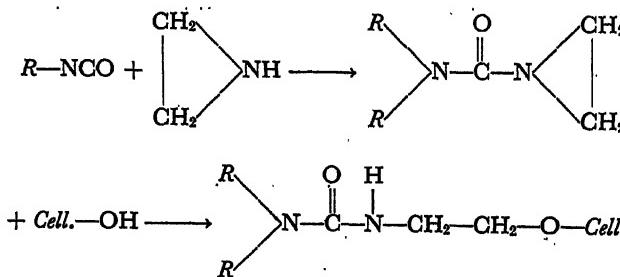
The ethylene ureas, however, are particularly valuable for rendering cellulose hydrophobic or even completely water-repellent by the use of octadecy-

lene urea.* Treatment of viscose- or acetate-rayon filaments with reaction products of aliphatic, cycloaliphatic and arylaliphatic isocyanates with ethylenimine gives acid-dyeing affinity.[19] A compound of this nature, diethylene urea 1,6-hexamethylenediamine, has been applied to cuprammonium regenerated rayon, dried, and cured, to control the swelling of the fiber and impart a hydrophobizing effect with consequent shrinkage control.[20]

Synthetic resins can also be applied to the rayon spinning solutions to change the chemical characteristics of the fiber. The resin can be added in soluble form, polymerized, or it can be formed in situ by addition of the reactive components. The use of resins always requires a curing stage at high temperatures (300°F), in like manner to that necessary on fabrics. This is a rather complicated procedure, however, for the filament stage. "Rayolanda" WS, recently produced on a commercial scale by Courtaulds, is a fully animalized viscose fiber obtained by addition of synthetic resins derived from cyanamide.

Many uses of resins are made possible during printing of fabrics. Besides the resins used as binders for pigments, it is often convenient to add the resin precondensates to the printing pastes containing acid dyestuffs and insolubilize by utilizing the heat of the drying and aging processes. Many urea derivatives, thiourea, dicyandiamide, biuret, guanylurea and guanidine in conjunction with formaldehyde have been suggested [21] for animalizing. Resin reactions carried out with different catalysts, acid or alkaline (tartaric acid or triethanolamine), impart affinity either for basic or acid dyestuffs. Good affinity for basic dyes has been obtained by using methylolurea, formaldehyde, and guanidine.

A new process to impart animalization consists in treating rayon with a solution of cyanamide, hydroxyaldehyde or an aliphatic hydroxyketone at a pH of 8.[22] Polymerization products derived from reaction of amines with acylated halogen derivatives of polyvinyl acetate can be used.[23]

Animalization Methods Employing Mixtures of Proteins and Other N-Bearing Compounds. Albuminoids can be used in conjunction with resins and many of the other N-bearing products mentioned, to give simultaneous acid-dyeing properties and greater tensile strength, thus controlling the tendency of the proteinized cellulose to swell excessively. The use of aerating compounds gives rayons of a tubular cross section with a multitude of entrapped air spaces which lighten the fiber and also serve to render it more highly animalized because of greater heat-retention qualities. Sulfur-containing compounds such as polysulfides can be used in place of those containing nitrogen.

Ethylenimine can be used with casein to treat the coagulated rayon, then subsequent treatment with a solution of phenylisocyanate in cyclo-

* Persistol VS of I. G. Farbenindustrie A.-G.

hexanol. Instead of using the cyclohexanol-surface method, the three reaction products have also been added directly to the viscose mass to produce the German "Vistralan." This fiber was one of the first successful types and had a soft wool-like feel and appearance, but lacked tensile strength. Changes have been made to this modified rayon by adding an amine condensate * of a methylated trichloroparaffin with polyethylenemine.

French Patent 646,529 refers to a treatment for viscose fabrics which imparts affinity for acid dyes together with anti-crease effect, by impregnating with a mixture of formaldehyde and isocyclic sulfonic acid (e.g., o-aminophenol sulfonic or naphthenic acids). Condensation products of cyanamide-formaldehyde impart to cellulose (and also to animal fibers) increased affinity for dyestuffs together with improved washfastness of direct colors.

The field of dye-fixing is being actively investigated and it is already possible to make some inexpensive direct dyes perform the function of more complicated aftertreated (copper, formaldehyde or diazo) colors by using dye-fixing agents based on resin reactions. At the present time, the entire gamut of dyestuffs of any fastness requirement can be readily made available; nevertheless, dye-fixing appears to have a promising future.

Synthetic Fibers Other Than Cellulosic

The entire program of animalization and chemical modification of cellulose may be undermined considerably by the efforts being made to develop new fibers to supplement the present cellulosic synthetic rayons, cotton, and even wool.

Nylon, a linear polyamide prepared by polymerization of the polymethylene amide of adipic acid, has a structure closely resembling protein. This fiber is produced by extrusion in the molten state and is thus a true plastic. The resultant textile fiber is lustrous, strong and highly elastic, moth and mildewproof. In many respects it is superior, for some uses, to any natural or synthetic fiber. Nylon can be dyed with acid, metallized chrome, mordant, direct, acetate, and Vat dyes, and it absorbs them in like manner to wool and silk. At first appearance, it would seem that the studies of cellulose animalization should be discontinued and that nylon fills the requirements. A second glance, however, showing the dependence of nylon on phenol as a starting material, indicates the fallacy of such a conclusion.

Vinyon, a copolymer of vinylchloride and vinylacetate, is a new plastic fiber spun like acetate rayon, which has the unique properties of being

---

* Solidogen BSE of General Dyestuff Corporation, New York.

water-repellent and flameproof and, therefore, is ideal for certain uses. It is, however, difficult to dye, and at present it is being confined to special uses (industrial filters). Saran, Orlon and Terylene are too new to show sufficient indication of potential uses.

Animal Fibers

Silk is a natural protein providing fibers of distinctive feel, luster and draping qualities. Considering its additional features of excellent wearing and heat-insulating properties, it is in many respects an almost perfect fiber. However, it is so high in cost that it suffers in economic comparison with those synthetic fibers closely resembling it in appearance. Nylon, in fact, has actually proven itself superior in practically all major fields. Silk is steadily losing its position today, even as a luxury item.

Wool, although third in the trio of fibers—cotton, rayon and wool— maintains a very important position and remains undisputed for warm clothing in the more temperate zones of the world. It is being continually subjected to competition in many fields of clothing, particularly for dress wear and summer suitings and in the latter is losing ground to the resin-treated spun-rayon and spun-acetate fabrics. The wool fiber is extremely complex structurally and possesses most of the necessary requisites for the production of satisfactory fabrics. However, wool as a protein, and, therefore, unrelated to any of the cellulosics, could not, until very recently, take appropriate advantage of the new developments in rayon and plastics technology.

The wool fiber, according to Speakman and Astbury, consists of long peptide chains bridged by cystine and salt linkages. It possesses amphoteric properties (isoelectric point—4.5) allowing it to function both as an alkali or as an acid and is, therefore, extremely reactive chemically. It dyes with basic, acid, chrome, direct, vat, and solubilized-vat leuco-ester dyestuffs, thus permitting great flexibility of choice to dyers and colorists in both shade and fastness properties. It possesses the property of felting which gives great ease of manipulation during the finishing of its fabrics. Finishing methods based on this characteristic have been the keynote of wool-fabric processing for centuries.

The necessity of chemical and physical modification of wool, other than dyeing, has not been considered until recent years. The felting character of wool, allowing it to mat and give compact fabric structure at the same time, causes objectionable shrinkage during mild washing of the garments. Another serious defect of the wool fiber, the destruction by moths, causes considerable economic damage annually. Unfortunately, the wool merchants' attitude, "the moth is our best customer," continues

to prevail and places the entire wool industry at a disadvantage in comparison to the low-cost spun-synthetic fibers and to nylon.

Permanent finishing of wool for men's wear has not created a serious problem because the garments are usually cleaned with solvents and, therefore, no shrinkage difficulties arise. This problem is extremely evident, however, in articles such as socks, underwear, sweaters, etc., which are washed frequently and rendered unwearable. This waste places an additional burden on the economy of wool fabrics. Methods now exist, to control both the ravages of moths and the shrinkage of woolen fabrics. Their adoption by the wool industry will assist in helping it to maintain its most suitable markets and will also create the necessary impetus to further technical investigation in applying available basic research to wool.

Mothproofing. Control of the damage made by moths need not be left to the discretion of the housewife, imposing an uneconomical drudgery, but can be easily accomplished by the manufacturer by one of the following methods:

(1) Use of moth repellents based on fluorides or silicofluorides that render the wool completely resistant to moths. These salts are of low cost, water-soluble and easy to apply. They are fast to dry cleaning, but not to water or mild washing.

(2) Use of quaternary ammonium and phosphonium complexes. These are water-soluble, colorless, and possess affinity for the wool fiber. They impart good fastness to dry cleaning and to water, but are not fast to washing. One type,[24] 3,4-dichlorobenzyl triphenyl phosphonium chloride, can be applied in the presence of nonionic detergents to give simultaneous mothproofing and cleansing.

(3) Durable mothproofing treatment can easily be obtained by the use of colorless acid dyestuffs which possess excellent affinity for wool in a boiling acid bath over a regular one-hour dyeing cycle. One type,[25] pentachlor dihydroxy triphenyl methane sulfonate, can be applied during the original dyeing of the fiber in any form or machine, but it is not as cheap or simple to use as the first two methods.

A fiber-modification approach to the moth problem has been undertaken by Harris,[26] based on a series of researches which demonstrated that the chemical structure of the wool fiber could be altered in such a way as to affect its chemical, biological, and mechanical properties profoundly. A review of this research indicates that the outstanding property of wool is its elastic recovery based on the fact that wool, like rubber, is composed of long, flexible molecular chains that tend to remain in folded configurations until subjected to stretch, but readily return to the original positions once the strain is removed. This unusual property of wool is the result of the presence of covalent disulfide link-

ages between the main chains which strengthen the fiber, mainly in the wet state. These cross links are easily destroyed and account for the sensitivity of wool to alkalis, oxidizing and reducing agents as well as its edibility by moths. Logically, it would appear that substitution of these cross links with more stable ones would enhance the stability of wool and eliminate its undesirable qualities:

By reduction of the disulfide linkages with thioglycollic acid, followed by rebuilding with trimethylene bromide, Harris obtained some of the desired results:

$$W—S—S—W + 2HSCH_2COOH \longrightarrow 2WSH + (SCH_2COOH)_2$$
$$2WSH + Br(CH_2)_3Br \longrightarrow W—S—(CH_2)_3—S—W + 2HBr$$
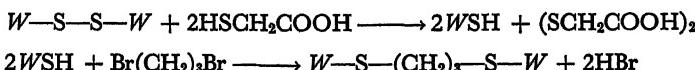
Use of cheaper materials, such as sodium hydrosulfite in conjunction with sodium sulfoxalate, $NaHSO_2 \cdot CH_2O$, to give $W—S—CH_2—S—W$, provides a similar linkage and points to the way of future research.

Other lines of chemical modifications of wool which change the entire physical characteristics as well, are those based on dissolving proteins similar to wool, i.e., feathers, and spinning them according to the principles of synthetic fiber extrusion. Dyeing affinity of wool towards acid dyestuffs can be changed by treatment with chlorine, which imparts greater affinity for acid dyestuffs. Chlorinated and unchlorinated wool dyed together give two-tone effects. Wool may be made partially resistant when treated with tannin to reduce the affinity for acid dyestuffs. Sulfuretted phenols are used in union dyeing of cellulosic fibers and wool with direct colors to keep the dye from staining the wool. This appears to be based on the fact that sulfuretted phenols block the amino dye-attracting groups of the fiber.

Wool can be dyed cold in the presence of a large amount of acetic acid. This method of dyeing could have greater application if the wool industry made more use of continuous-dyeing machinery, as is common in the processing of cotton and rayon. Present methods of wool dyeing require long time factors and it is felt that a study of the potentials would greatly curtail time employed in dyeing.

Quaternary ammonium compounds on the order of octadecyl pyridinium bromide have been used to impart greater affinity to wool for acid dyestuffs, but oftentimes such a method leads to difficulties because of the possibility of precipitation of dye specks as a result of the reaction between the two components, an anionic dye and a cationic treating agent.

Shrinkage Control of Wool Fibers. Shrinkage control has been practiced for many years by the use of chlorine. New chlorine-containing chemicals based on sulfuryl chloride, chloro sulfonic acid, and brominated products have recently been proposed. All these methods, except

those employing hypochlorite, are of minor importance. Those utilizing hypochlorite have received added attention since World War II, but they are difficult to control. Chlorine reduces the shrinkage of wool by partially dissolving the fiber surface and by equalizing the coefficients of friction. The danger of attacking the fiber severely calls for rigid conditions of reaction. These difficulties led to the use of synthetic resins.

Resinification of the Wool Fiber. Speakman[27] suggested the reduction of felting power of wool by changing its elastic properties. Resins such as urea-formaldehyde and also the thermoplastic types, acrylates, vinyls, and styrenes were found inadequate until it was shown [28] that alkylated methylol melamine resins can be successfully applied to give good control without affecting the hand of the wool. Melamine resins are obtained from cyanamide by polymerization to form first dicyandiamide and then melamine (triamino triazine). Melamine reacts with formaldehyde and then by methylation of the product the alkylated monomer is formed. The latter is completely water-soluble and applicable to wool in the presence of a catalyst. When dried and cured, the resulting water-insoluble resin polymer is deposited inside and outside the wool fiber to give dimensional stability and thus control of shrinkage in fabrics so processed. This method is now in successful operation on a large scale.

Conclusion

An attempt has been made to telescope the trends of research and the recently awakened interest in the textile industry. The entire development of improved textiles is of such a fundamental nature and it overlaps so many other industries that it was felt advisable by the author to survey and coordinate the various technical ideas involved in the recent advancements of fiber modification. It is sincerely hoped that this paper may be of interest mainly in this connection.

REFERENCES

1. Kosslinger, K. (to I. G. Farbenindustrie), U.S. Patent 2,200,452 (1940).
2. Mark, H., *Am. Dyestuff Reptr.*, **36**, 323 (1947).
3. Baldwin, A. (to Imperial Chemical Industries Ltd.), U.S. Patent 2,146,392 (1939).
4. Foulds, R. (to Tootal Broadhurst Lee Co., Ltd.), U.S. Patent 1,734,516 (1929).
5. Lippert, A. (to Joseph Bancroft and Sons Co.), U.S. Patent 2,103,293 (1937), U.S. Patent 2,148,316 (1939).
6. Borghetty, H. C. (to Aspinook Co.), U.S. Patent 2,050,156 (1936); U.S. Patent 2,196,655 (1940).
7. Swan, T. (to Cluett, Peabody & Co.), Brit. Patent 473,478 (1937).
8. The Calico Printers' Association Ltd., Germ. Patent 670,471 (1939).
9. Sandoz Chemical Works, Inc., Brit. Patent 464,921 (1937).
10. Munz, F. (to General Aniline Works, Inc.), U.S. Patent 2,080,543 (1937).
11. Olpin, H. (to British Celanese Ltd.), U.S. Patent, 2,403,019 (1946).

12. Tagliani, G., *Melliand Textilber.*, **7**, 765 (1926).
13. D'Ambrosio, A., Fr. Patent 830,101 (1938).
14. I. G. Farbenindustrie, A. G., Brit. Patent 440,449 (1935); Brit. Patent 440,647 (1935).
15. Bloxam, A. (to Ciba Co.), Brit. Patent 370,860 (1932).
16. Burke, W. (to E. I. duPont de Nemours & Co.), U.S. Patent 2,391,942 (1946).
17. I. G. Farbenindustrie, A. G., Fr. Patent 807,984 (1937).
18. I. G. Farbenindustrie, A. G., Fr. Patent 809,377 (1937).
19. Bestian, H. (to General Aniline & Film Corp.), U.S. Patent 2,317,965 (1943).
20. F. I. A. T., **41**, Great Britain Intelligence Objectives Subcommittee.
21. Sisley, J. P. to Société Anonyme des Matiers Colorantes et Produits Chimiques), Fr. Patent 713,283 (1931).
22. Cameron, W. (to Courtaulds Ltd.), U.S. Patent 2,318,464 (1943).
23. Schlack, P. (to Walther H. Duisberg), U.S. Patent 2,267,842 (1941).
24. Lommel, W. (to I. G. Farbenindustrie A.-G.), U.S. Patent 1,921,364 (1933).
25. Weiler, M. (to I. G. Farbenindustrie A.-G.), U.S. Patent 1,948,894 (1934).
26. Harris, M. and Brown, A. E., *Am. Dyestuff Reptr.*, **36**, 316 (June 16, 1947).
27. Speakman, J. B., "Chemistry of Wool and Related Fibers," *J. Textile Inst.*, **32**, 83 (July, 1941).
28. Johnstone, E., Jr. (to American Cyanamid Co.), U.S. Patent 2,329,622 (1943).

WOOD—AN EXAMPLE OF A
COLLOID SYSTEM

Harold Tarkow and Alfred J. Stamm

Forest Products Lab., Madison, Wisc., Forest Service,
U. S. Dept. of Agriculture

WOOD CAN BE STUDIED from the point of view of the functions of the various components in the living tree or from the point of view of the manner in which these components influence the properties of the commercially used material. This review adopts the latter approach.

Description

Wood is a heterogeneous substance in both the chemical and physical sense. What may appear to the unaided eye to be a chemically-uniform material is on a microscopic or submicroscopic scale a mixture or loose chemical combination of cellulose, pentosans, polyuronides, pectins, hexosans, lignin, tannins, gums, starch, and ash-yielding inorganic material. Physically, heterogeneity is present at all levels of observation from that of the naked eye down to that of the x-ray beam. Variations in the amounts of these different components and in the complexity of the cellular system account for the recognition of more than 20,000 woody species in the world.

The common species from which the woods of commerce are derived, are divided into two classes, based on their manner of producing seeds. These are the gymnosperms and angiosperms, that is, the softwoods and hardwoods of common usage. Paradoxically, however, the distinguishing features of the woods from these two classes is not their strength; rather, it is primarily their structure and secondarily their chemical composition.

Softwoods are characterized by possessing a relatively simple microscopic cell structure. The most prominent cell type is the tracheid—a radially-arranged, vertically-directed (i.e., in the tree direction), completely enclosed cell ranging in length from 2 to 7 millimeters and in

diameter from 20 to 60 microns. These cells are separated from and bonded to one another by the material of the middle lamella—a resinous substance believed to be composed essentially of lignin.[31] The total volume of the tracheids constitutes more than 90 per cent of the volume of softwoods.[21] The remaining volume is composed, essentially, of thin-walled, short, radially-directed cells, collectively making up the medullary or wood rays, together with still smaller volumes of vertical and horizontal resin ducts.

In sharp contrast, hardwoods have a much more complex capillary structure. They have many more specialized cell types. Some hardwoods, in fact, possess as many as six different types of cells. The most prominent of these from the point of view of visibility are the vessels or pores, which are radially arranged and vertically directed in the wood. They range in diameter from 25 to 500 microns and constitute between 15 and 30 per cent of the volume of hardwoods.[21] Unlike tracheids, they do not seem to be completely enclosed, but appear to be tubes of indefinite length. In some species (red oak), the tubes are clear. One can readily blow air through a stick of red oak several feet long. In others (white oak), the tubes are clogged with resinous material, called tyloses, which make the wood less pervious. Also vertically-directed, but completely enclosed, are the wood fibers, ranging in length from less than a millimeter to about 2 millimeters, and in diameter from several microns to about 20 microns. They constitute 25 to 75 per cent of the volume of hardwoods.[21] Hardwoods, too, have a system of medullary rays. These, together with other less prominent cell types, make up the remainder of the capillary structure of hardwoods.

During growth of the tree, the thickening of cell walls gives weight and strength to the wood. The extent of thickening increases with the lateness of the growing season. It is thinnest in those cells formed in the spring and thickest in those formed in the summer towards the end of the growing season. With the resumption of growth in the spring, the accompanying formation of thin-walled cells, adjacent to and concentric with the thick-walled cells of the previous summer's growth, delineates the annual ring.* In general, the strength properties of wood increase as the proportion of thick-walled cells increases, that is, as the density of the wood increases.

The cell walls contain numerous circular areas of a radius of 20 to 50 microns, where thickening has not occurred. The only barrier between adjacent cells in these areas is the middle lamella. These areas are known as pits. The membranous middle lamellae, called pit membranes, are

---

*The thickness of annual rings records annual growing conditions. From the rings of living California Sequoias Prof. A. E. Douglass charted weather conditions as far back as 3,000 years. See paper by C. A. Reeds in Vol. V of this series.—*Ed.*

permeable to water vapor.[45] Furthermore, owing to the presence of punctures, cracks, and other permanent openings [58] whose radii are of the order of 30 microns,[58] they are often permeable to liquids, except when clogged by resinous materials. These pits are especially numerous in softwoods and, unless clogged, offer means for the transfer of liquids from cell to cell.

The cell walls are thickened by the deposition of material on the original, extensible primary wall. The deposition is not a random one,

Figure 1. Softwood fiber showing intact wrapping and limited swelling in that zone and the high degree of swelling or ballooning where the wrapping is removed.

but occurs in concentric laminae. Each lamina contains cellulose molecules that, in certain areas, are oriented parallel to each other to form what Nageli called micelles and what Meyer [26] more correctly called fringed micelles or crystallites. These areas of regular orientation are responsible for the x-ray diffraction pattern of wood.[4] The cellulose chains outside these organized areas, together with the other carbohydrate constituents and lignin, constitute the amorphous regions of wood. These regions are relatively more accessible to chemicals than are the regions of the oriented cellulose chains. Recent studies on the saccharification of wood [35] show that the carbohydrates here undergo hydrolysis at a much faster rate than do the cellulose molecules in the organized regions. The more rapid rate of hydrolysis of the carbohydrates in the amorphous regions may be due in part, however, to the inherently greater hydrolysis rate of 1, 4 xylo pyranosidic linkages than that of the gluco pyranosidic bonds [36] that are present in the organized areas. The

amorphous regions acetylate with much greater ease than do the crystal-
lite regions,[60] and they are largely responsible for the hygroscopicity of
wood and very likely play an important role in contributing to wood's
toughness, flexibility, and steam bendability. The fringed micelles are
in turn oriented. The direction of this orientation varies in the different
laminae of the cell wall. Bailey and Kerr [23] have studied the micellar
orientation in wood by optical means. They report the existence of at
least three concentric zones in the cell-wall thickening: a thin zone (one
micron), just inside the primary wall, in which the micelles are oriented

Figure 2. Spiral windings in leaves
of Nerine fothergillii.

approximately at right angles or at some large angle to the long cell axis
(grain direction); a relatively thick, but variable, central zone, in which the
orientation is parallel or at some small angle to the cell axis; and, finally,
a thin layer similar to the first one, which is adjacent to the cell cavity
or lumen. Swelling studies of wood fibers by Ritter [32] have vividly shown
the presence of the first and central zones. Thus, in Figure 1, there is
shown the behavior of an isolated wood fiber in 60 per cent sulfuric acid.
In one area, the first layer either has been dissolved away or rolled back
so as to permit the thick central zone to swell extensively beyond the
normal water-swollen dimensions. Figure 2 shows some of the spiral
windings in the leaves of Nerine fothergillii.[59] It is perhaps an extreme
example of the orderly manner in which micelles are oriented in plant
tissues.

Lignin and ash-forming inorganic material interpenetrate the cellulose throughout the cell wall in a continuous manner. This has been strikingly shown by Ritter,[34] who removed different components from blocks of wood and obtained blocks of residual lignin, cellulose, and even ash, having the skeleton form of the original blocks.

The peculiar orientation of the micelles in the cell walls, with the predominant orientation (central zone) being approximately parallel to the grain direction (i.e., the longitudinal axis), bestows on wood highly anisotropic properties. Thus, the strength properties are much greater in the grain direction than in the transverse direction. The tensile strength may be 40 times greater, the crushing strength 7 times greater, and the modulus of elasticity 150 times greater. Swelling in the grain direction is relatively insignificant compared to that in the transverse direction. Although the coefficient of thermal expansion of wood is so small as to be unimportant for ordinary consideration, nevertheless, the linear expansion may be 10 times as great in the transverse direction as in the grain direction.[62]

In addition to this pronounced longitudinal-transverse anisotropy that is associated with the predominant longitudinal orientation of the micelles, there is a much smaller tangential-radial anisotropy. Thus, strength properties are generally slightly greater when measured radially (perpendicular to the grain and to the annual rings) than when measured tangentially (perpendicular to the grain and parallel with the annual rings). Radial swelling is generally one-half to two-thirds of the tangential swelling, and the linear coefficient of thermal expansion in the radial direction may be 25 per cent less than in the tangential direction. This phenomenon of tangential-radial anisotropy has been variously interpreted in terms of the influence of the radially-directed medullary rays,[9] and in terms of the slightly different orientation of the micelles on the radial walls than on the tangential walls.[4, 33]

Hardwoods and softwoods show some differences in their chemical composition. Perhaps the most striking difference is in the pentosan content. Hardwoods, on an average, contain about 20 per cent of pentosans and softwoods about 10 per cent. The lignin content of hardwoods varies between 20 and 25 per cent, and that for softwoods between 25 and 30 per cent. The constitution of lignin *in situ* is very uncertain. It is generally agreed that some of the side groupings on the lignin residue are different in the two classes of wood. Ethanolysis of softwoods yields guaiacyl derivatives, whereas hardwoods yield both guaiacyl and syringyl derivatives.[24] The cellulose isolated from softwoods and hardwoods have about the same molecular weight. There have been numerous measurements of the degree of polymerization of cellulose. The generally accepted belief at present is that in the native state it is

indefinitely high. Berkley [3] has obtained values as high as 10,500; however, he shows that even slight handling or weathering may reduce the degree of polymerization to the commonly measured values.

Adsorption Phenomena

The reaction of water with wood has been extensively studied. The interest in this reaction arises from the intimate association of water with the wood in the living tree and from the pronounced dependence of many properties of wood on the moisture content. Thus strength, dimensions, density, and electrical conductivity are dependent on the moisture content. As it comes from the tree, wood may contain as water, from one-third to three times the weight of the oven-dry material. Dry wood is very hygroscopic. Its adsorption isotherm, similar to that of water on cellulose,[1] is the typical sigmoid-shaped curve—moderate adsorption at very low relative humidities, pronounced adsorption at high humidities, and relatively less adsorption at intermediate humidities. The first few percentages of water are adsorbed with the evolution of much heat which corresponds to a large decrease in entropy. The decrease in entropy, here, is approximately equal to that for the formation of ice.[48] This suggests that the initially adsorbed water is held fixed and oriented in a manner similar to the water molecules in ice. Green wood gives the characteristic x-ray diffraction pattern of crystalline cellulose; thus adsorption must occur only in the amorphous regions. The initially adsorbed water is very likely hydrogen-bonded by the hydroxyl groups in these regions. Pauling[29] has similarly interpreted the initial adsorption of water by proteins. His calculations show that adsorption occurs to the extent of one molecule of water per polar side chain capable of forming hydrogen bonds. Hermans[20] has suggested the formation of hydrates of cellulose at the low humidities. His analysis, based on the formation of hydrates, leads to an expression that reduces to the Langmuir adsorption equation. Hermans points out the similarity between the adsorption isotherm of water on cellulose and that of water on sulfuric acid. In the latter case, it is generally believed that hydrates do form at low relative humidities.

There is no generally accepted theory that accounts for the adsorption of water by wood or cellulose at relative humidities above 30 per cent. Since the differential heat of condensation and the entropy decrease and eventually approach those values for the condensation of water vapor, it has been suggested that capillary condensation is occurring.[43]

The relationship between the radius of a capillary (r) and the relative vapor pressure of water in the capillary (p/p_o) is given by the Kelvin equation:

$$r = - \frac{2\sigma M}{d_0 RT \ln \frac{p}{p_0}}$$

where σ is the surface tension of the adsorbed liquid, M is its molecular weight, d_0 is its density, R is the gas constant and T is the absolute temperature. None of the microscopic capillaries described in the previous section is small enough to reduce the relative vapor pressure of water below 0.95, so that condensation in them below this relative vapor pressure cannot occur. The moisture must be adsorbed in the cell walls. In doing so, the moisture creates its own capillary system, which Stamm [43] calls transient capillaries. That this moisture, adsorbed above 30 per cent relative humidity, is not surface-bound water, is suggested by the fact that although the first few per cent of adsorbed water has no solvent properties, the water adsorbed beyond 30 per cent relative humidity seems to have normal solvent properties.[40] This has been refuted by Barkas,[5] who claims that as much as 20 per cent of the adsorbed water has no solvent properties and that the remainder is water of capillary condensation. There is some doubt, however, whether Barkas measured equilibrium concentrations.[47]

The question whether the moisture other than the surface-bound water is adsorbed through the mechanism of capillary condensation is still very much debated.[16, 25] The accompanying volume change during adsorption has been attributed to the wedging apart of the swelling units in wood by the capillary-condensed water. Thermodynamically, this is unsound. The existence of tensile forces in water in submicroscopic capillaries would tend to shrink rather than swell the flexible cell walls. Haines and McIntosh claim to have observed such a shrinkage of carbon rods and have interpreted it to mean the presence of air-water interfaces in the rods.[18]

The adsorption of water by wood may be caused by the tendency of shrinkage stresses to relieve themselves.[14] Wood is formed in the presence of liquid water and in its natural state it is stress-free. On exposing green wood to some relative humidity below 100 per cent, desorption occurs. When the moisture leaves the transient capillaries in the cell walls, air-water interfaces are created. The remaining moisture is under a tensile stress whose magnitude is some inverse function of the relative humidity. This tensile stress tends to draw the walls of the transient capillaries together. The resulting strain in the cell walls is an elastic one and has associated with it a stress acting in a direction opposite to that of the tensile stress in the capillary water. At equilibrium the two stresses are equal. On removing another increment of water, by lowering the relative humidity, the resulting increase in tensile stress in the capillary water

produces an increased elastic strain and stress in the cell walls. A new equilibrium is established. Thus, at all moisture contents below saturation, shrinkage stresses exist, and these increase in magnitude as the moisture content decreases. On reexposing the wood to a higher relative humidity, the principle of Le Chatelier dictates that water be adsorbed, for, in so doing, the shrinkage stress is reduced.

Expressed differently, the shrinkage stress resulting from the elastic strain in the cell walls may be considered to be a "solution" pressure analogous to the solution pressure of sulfuric acid in water. The pressure in the latter is due to the kinetic energy of the acid molecules and decreases with increasing moisture content. Likewise, the "solution" pressure in the cell walls is due to the kinetic energy of the swelling units, and it, too, decreases with increasing moisture content. In fact, Hermans [20] attributes the portion of the isotherm beyond the first inflection point to a type of solution of cellulose hydrate in water. His analysis, based on the existence of an ideal solution, shows that a pronounced increase in adsorption should occur at high relative humidities. His argument has qualitative validity; quantitatively, however, it is insufficient, for ideality of solution certainly does not obtain, any more than it does in sulfuric-acid-water solutions.

Perhaps the most successful of the general adsorption theories, at least for rigid, nonswelling capillary systems, is that of Brunauer, Emett, and Teller,[10] based on the assumption of multimolecular adsorption. It is an extension of the Langmuir theory of surface adsorption with continued adsorption on successive but incomplete layers. The theory predicts the heat of adsorption accompanying the adsorption of the first layer and the specific surface area of the adsorbent. Babbitt [1] has applied the theory to the adsorption of water by wood and has calculated from it the net heat of adsorption; that is, the heat of adsorption onto the bare surface minus the normal heat of condensation. Although he reports excellent agreement with the experimentally-determined value, a reconsideration of his data has shown the agreement to be very poor.[2] The theory assumes the presence of a definite amount of surface area that remains constant throughout the entire adsorption. This assumption is not valid for wood, for at zero per cent relative humidity, the specific surface area is virtually zero. (Actually, it is the surface area of the microscopic capillaries, $2 \times 10^3 cm^2$ per gram.[50]) As the moisture content increases, the specific surface increases.[46]

Swelling of Wood

Regardless of the mechanism whereby water is adsorbed by wood, such adsorption (or desorption) is accompanied by volume changes. At low

relative humidities (less than 30 per cent) the volumetric swelling of wood is less than the volume of water adsorbed. It is generally believed that the strong surface adsorption occurring at low humidities results in the compression of the water.[15] Stamm [44] has calculated the specific volume of the initially adsorbed water to be 0.8 cubic centimeter per gram. The specific volume increases rapidly towards unity with increasing adsorption. Hermans,[20] on the other hand, does not accept this theory because of the very large compressive forces required to decrease the specific volume of water. Instead, he assumes that the cell walls contain submicroscopic capillaries or voids of relatively small total volume into which moisture can enter without contributing to the overall volume. Stamm and Hansen,[46] however, have shown that the density of cellulose as determined by helium displacement agrees very well with the value calculated from Meyer's and Misches' [27] most recent x-ray diffraction data of crystalline cellulose. They question whether submicroscopic voids do exist.

Between 30 and 95 per cent relative humidity, the adsorbed water can be considered, as a first approximation, to add its normal volume to that of the wood. At the higher humidity, not only do the cell walls continue to adsorb water and increase in volume correspondingly, but now the permanent capillaries of sufficiently small size (10 mu) adsorb water, since they are capable of holding water at reduced relative vapor pressures. At still higher humidities, water continues to condense in the tapered ends of the cell cavities and eventually in the main body of the cavities. Since the cavities constitute a large fraction of the total volume of the wood (wood of density 0.4 gram per cubic centimeter has a fractional void volume of 0.75), it is understandable why large amounts of water are adsorbed at these very high humidities for small increments in relative humidity.

Since at relative humidities below 95 per cent, adsorption occurs only in the cell walls, extrapolation of the adsorption isotherm to 100 per cent humidity gives approximately the moisture content of the cell walls at saturation; that is, it gives the moisture content of wood at saturation with no contribution from that condensing in the cell cavities. This extrapolated moisture content is called the fiber saturation point. Its value is approximately constant for all species, namely 26 to 30 per cent, although certain woods having high extractives contents (for example, redwood) have low fiber saturation points. If these extractives are leached out, the wood then acquires the normal fiber saturation point.[48] This saturation moisture content is an important constant, for many properties of wood are dependent on the moisture content up to the fiber saturation point. With increasing moisture content up to the fiber saturation point, strength properties decrease and then become constant. The

logarithm of the electrical conductance increases linearly with moisture content up to the fiber saturation point, above which the rate of increase is appreciably smaller.[39]

Barkas has questioned the fundamental significance of the fiber saturation point,[6] since at relative vapor pressures slightly less than unity, water condenses in the cell cavities. Furthermore, he reports he has observed shrinkage in wood at moisture contents as high as 100 per cent. He also reports that strength continues to decrease at moisture contents above the fiber saturation point. The fact that water condenses in the coarse capillaries above 95 per cent relative humidity is recognized; for that reason the extrapolation is made from below this relative humidity, disregarding the experimentally-observed course of the curve at higher humidities. The observance of shrinkage at moisture contents as high as 100 per cent is undoubtedly real. Others have observed this [48] and have attributed it to surface shrinkage stresses resulting from the moisture content at the surface dropping below the fiber saturation point, although the overall moisture content is above the fiber saturation point. There is still another experimental complication that may bring about shrinkage at moisture contents in excess of 30 per cent. Barkas used cross sections of beech, 1 millimeter thick, believing thereby to have cut open each wood fiber. The average length of a wood fiber is 1 millimeter; however, there is a certain fraction of the total number that is smaller in size. Many of these, including the still smaller parenchyma cells, will not have been cut open. Water in them will not tend to evaporate until the relative humidity falls below that in equilibrium with the water held in the fine pit membrane pores with radii ranging from 10 to 40 mμ, which, according to the Kelvin equation, occurs only at about 95 per cent relative humidity. Between 100 and 95 per cent humidity, however, some moisture will have left the cell walls. Thus, at 95 per cent humidity some shrinkage will have occurred, although the overall moisture content is appreciably in excess of 30 per cent. The fact that Barkas reports decreasing unit strengths at moisture contents above the fiber saturation points is due to the peculiar way in which the unit strengths are calculated. He contends that when a compressive load is applied to a unit area of wood, the load is resisted not by the unit area but by the fractional cross-sectional area of the wood-water aggregate in this unit area of wood; hence the strength values he reports are higher than the conventional values. From a fundamental point of view, this is perfectly reasonable, provided the moisture content is less than 30 per cent. At moisture contents in excess of the fiber saturation point, he assumes that the moisture in the partly filled cavities continues to support part of the load. This is questionable. The swelling pressure above the fiber saturation point is virtually zero; at any rate it

is less than the applied load and hence cannot contribute to the support of the load. Since he assumes that with increasing moisture content above the fiber saturation point, the area supporting the load continues to increase, obviously the calculated unit load values must decrease.

The approximate linear relationship between volume of adsorbed water and the external swelling suggests that the cell cavities remain constant in size during swelling and shrinking. This may be surprising, for one would expect the cavities to increase in size when wood swells just as a hole cut in a sheet of rubber expands when the sheet is stretched. Stamm has shown [41] through permeability studies that the cavities do remain approximately constant in size between zero and 95 per cent humidity. Recent microscopic examination at the U.S. Forest Products Laboratory has confirmed this for the relative humidity range of 20 to 90 per cent. Assuming the cavities do remain constant in size and that shrinkage is due only to the removal of water from the cell walls, it can be shown that the volumetric shrinkage from the green to the oven-dry condition is given by:

$$S = FD$$

for thin stress-free cross sections, where S is the volumetric shrinkage, F is the fiber saturation point expressed as volume of water per gram of wood, and D is the oven-dry-weight-green-volume density. The observed linearity of the shrinkage-density plot [42] is added evidence of the constancy of the cell cavities during swelling and shrinking.

The reason for this constancy in dimensions of the cell cavities during swelling is believed to be due to the nature of the inner layer adjacent to the lumen wall, which, owing to the orientation of its crystallites and owing to the incrustation of dead protoplasm on the surface, is able to withstand shrinkage stresses tending to distort the cell cavity. That the inner wall has a great resistance to internal buckling, is shown by the fact that although dilute aqueous sodium hydroxide swells wood beyond the water-swollen dimensions, the alkali concentration must exceed 6 per cent before internal buckling occurs.

Barkas [7] believes that the shrinkage that occurs between saturation and 85 per cent relative humidity is not caused directly by the removal of an equivalent volume of water from the cell walls but to the increased swelling pressure that he assumes can be calculated with the Katz swelling equation. The imposition of the swelling pressure, acting hydrostatically, produces a strain—the observed shrinkage. In fact, he contends that the ratio of the stress (swelling pressure) to the strain (volumetric shrinkage), which he calls the shrinkage modulus, is the same as the mechanical modulus of compression when wood is subjected hydrostatically to an external load. Shrinkage strain, however, is caused by a

contraction of the cell walls, with the cell cavities remaining constant in size and shape; whereas compression strain is due in part, at least, to bending, plastic strain, and partial collapse of the cell cavities. This is especially true at high moisture contents where the rigidity of the cell walls is low. Some contraction of the cell walls may occur. Barkas has shown [8] that the application of a compressive load on a small specimen of wood in equilibrium with 90 per cent relative humidity results in a slight reduction in equilibrium moisture content.

Dimension-Stabilization of Wood

The variation in moisture content with relative humidity is undoubtedly the outstanding defect of wood. Subsequent swelling and shrinking is very disturbing where close tolerances are necessary. Nonuniform moisture adsorption leads to warping and cupping. Rapid changes in humidity, such as occur when rainstorms are followed by bright dry weather, cause surface swelling and shrinking with subsequent checking and grain raising. Paint films may flake off, especially over the dense summerwood in which the volume changes are especially pronounced.

To moderate some of these effects, advantage may be taken of the smaller swelling of low-density woods; however, there is then a corresponding sacrifice in strength. Advantage may also be taken of the smaller radial swelling than tangential swelling. Wood for flooring, for example, is often quarter-sawn; that is, cut so that the width direction is parallel with the medullary rays. Swelling in this direction is less than would be the case if the wood were flat-sawn; that is, cut so that the width direction is perpendicular to the medullary rays. Paints, varnishes, and waxes have been used to reduce the effects of moisture. These are only partially successful. The presence of such films on wood does not alter the equilibrium swelling of wood. They simply reduce the rate of attainment of equilibrium.

The volume changes that occur in paper and cotton fabrics are more complex than those occurring in wood. On immersing paper or fabric in water, in addition to the swelling due to the entrance of moisture into the cell walls (which swelling may largely occur internally into the interfiber spaces with little contribution to the external volume), there is a further contribution to the external volume change owing to the partial dispersion of the fibers. With fabrics having a spun and woven construction, such pronounced interfiber dispersion may be reflected in an overall shrinkage of the fabric. This tendency to partial dispersion can be reduced by treating these materials with small quantities of resins (2 to 5 per cent). The reduction in interfiber dispersion can be looked upon

as being caused by the creation of a synthetic *middle lamella* between the fibers.

The dispersion of the fibers in wood, on the other hand, is impossible, owing to the bond (middle lamella) between the fibers. Small amounts of resin are without effect on the dimensional stability of wood.

The cell walls on swelling can accommodate a certain volume of foreign substance, or bulking agent. Under ordinary conditions, the maximum volume that can be accommodated is constant (about 30 cc per 100 grams of wood). In other words, wood can be considered to be a limited-swelling gel.[56] Since this volume is fixed, the amount of water that can be adsorbed by the bulked-out cell walls is less by an amount equal to the volume of bulking agent. The greater this volume, the smaller the volume of water that can be adsorbed. The substance deposited in the cell walls merely acts as an inert agent replacing an equal volume of normally adsorbable water. On treating wood with materials that can enter the cell walls, it is generally found that only a portion enters the walls. The amount entering is measured simply by the increase in oven-dry dimensions. The greater this increase, the greater is the volume of material in the cell walls and the smaller is the volume of water that can be adsorbed. With increasing molecular size, the amount entering the cell walls decreases. For this reason, waxes are without effect on the equilibrium moisture content. Specific interaction of the bulking agent with the wood is not necessary. Because of this, it is possible to calculate the effectiveness of a given treatment if the volume of material in the cell walls is known.[56] Acetyl groups in the cell walls (introduced by acetylation) are very effective in reducing the swelling and shrinking tendency of wood and seem to behave simply as inert bulking agents.[55] That is to say, the reduced swelling and hygroscopicity of acetylated wood is not directly caused by the replacement of hydroxyl groups; rather it is due to the physical bulking action of the acetyl groups. Thus equal volumes of acetyl and butyryl groups, when introduced into the same weight of wood, are equally effective in reducing the hygroscopicity and swelling, even though fewer hydroxyl groups are replaced in the butyrylated wood because of the larger molecular weight of the butyryl group.

Wood impregnated with certain phenolic resins to the extent of 25 to 30 per cent has very good dimensional stability.[51] The mechanism for this improved dimensional stability has been interpreted in terms of the above-described bulking action.[56] Barkas [8] claims it is caused by the imposition, by reaction, of a compressive load by the resin on the wood. The treated wood, known as *impreg,* has limited use, however, because of its reduced toughness. Resin-treated, compressed wood, known as *compreg,*[52] is commercially available. Its dimensional stability, high

strength, pleasing appearance, and ease of fabrication lend themselves highly for such specialty uses as dies, jigs, molds, gears, picker sticks of looms, and knife handles.

Certain treatments of wood reduce the total volume of foreign substance that can be accommodated in the cell walls; that is the total swelling due to the treating agent and subsequent adsorption of water is less than 30 cubic centimeters per 100 grams of wood.[57] Thus treatment of wood with acidified formaldehyde vapor reduces the subsequent swelling in water not by a bulking action of the formaldehyde, but rather by a cross linking or tying together of the swelling units in wood by methylene linkages. Severe embrittlement by the acid, however, places distinct limitations on the use of the material. That cross links between swelling units are present in formaldehyde-stabilized wood is shown by the fact that when such wood is immersed in concentrated-acid solution, the wood slowly swells and eventually attains the water-swollen dimensions of the untreated wood. Apparently, the aqueous acid hydrolyzes the acid-sensitive methylene linkages (cross links), and the structural units are no longer restrained from swelling. Fabrics have been stabilized with acidified aqueous formaldehyde. The stabilization here, however, is not caused by a reduction in the true hygroscopicity or in the amount of moisture adsorbed by the cell walls; rather, it is due to a reduction in the extent of interfiber dispersion. Thus, on treating cotton fabric with acidified formalin at 65°C, although the subsequent amount of water that can be imbibed may be reduced by 70 per cent, the true hygroscopicity is reduced by only 10 per cent.[28] Relatively dry formaldehyde vapor (acidified) must be used in order to reduce the true hygroscopicity.

Air-dry wood can be compressed to a density of about 1.30 grams per cubic centimeter. Pressures of the order of 2000 to 3000 pounds per square inch are necessary to accomplish this. The resulting material has relatively little void space. It has elastic strains in it, as shown by the fact that the wood slowly springs back to its uncompressed dimension. This recovery is hastened if the wood is moistened. Such compressed wood, then, has extremely poor dimensional stability, for the volume change on moistening is due not only to the adsorption of water but also to the springback. In fact, the contribution to the volume change by the latter may be twice the contribution owing to mere adsorption. If wood is compressed at a moisture content of about 9 per cent and at a temperature of about 160°C, the same densification can be obtained at a lower pressure, i.e., at 1500 to 2000 pounds per square inch.[54] Furthermore, under these conditions the compression is plastic rather than elastic. The compressed wood shows no tendency to recover its original dimensions even after moistening. The only volume change is that owing to the adsorption of water. This, however, occurs extremely slowly, since

compression has destroyed the coarse capillary passages through which most of the moisture diffusion normally occurs. This compressed wood, known as *staypak,* has an apparently good dimensional stability, and owing to the densification, has good mechanical properties.

Movement of Material Through Wood

Consideration has been given so far only to the equilibrium conditions in the wood-water system. The dynamics of the system and, in general, the movement of different materials into, through, and out of wood are of equal importance. The flow of materials in wood is involved in many industrial processes:

(1) *Seasoning of Wood.* A freshly cut log of 16 feet long by 1½ feet in diameter may have a liquid content in excess of 100 gallons. Before the wood can be used for general structural purposes, the moisture content must be reduced below the fiber saturation point and preferably to the average value it would come to under the conditions of its use. The reduction in moisture content is accomplished by air-seasoning or more rapidly by kiln-drying.

(2) *Treatment with Preservatives.* Wood that is used for outdoor purposes may frequently acquire a moisture content in excess of 20 per cent. Under these conditions, the wood is extremely susceptible to attack by fungi and molds or, when used under water, by marine organisms. Such attack may be reduced and the life of the wood prolonged by treating with certain chemicals. This treatment may consist of soaking or of high-pressure permeation so as to introduce toxic salts, such as zinc chloride, copper sulfate, mercuric chloride, and arsenicals or organic materials, such as pentachlorophenol or coal-tar creosote. In 1946, more than 300-million cubic feet of lumber, consisting mainly of railroad ties, farm fence posts, and telephone poles, were treated.

(3) *Treatment with Fire Retardants.* Certain salts, such as soluble borates and phosphates, impart fire resistance to wood. These salts are introduced by soaking or high-pressure permeation. In 1946, slightly less than 400,000 cubic feet of lumber were treated in this manner. In 1943, during the wartime peak of construction of airplane hangars, more than 5-million cubic feet were treated.

(4) *Pulping of Wood.* Wood is pulped to obtain the cellulose by permeating chips with certain liquors that dissolve out the lignin. The solubilized lignin then diffuses out of the chips.

(5) *Dimension-Stabilizing Treatments.* The stabilizing of wood consists in introducing certain chemicals into the cell walls. This is accomplished by soaking, high-pressure permeation, or exposing the wood to vapors.

In these different treating processes, three types of flow are recognized: capillarity, pressure permeability, and diffusion.

Capillary flow or capillarity occurs in response to the reduction in relative vapor pressure of the liquid in the capillary. The extent of penetration depends on the degree of wetting of the capillary wall, the diameter of the capillary, and the surface tension of the liquid. The degree of wetting is measured by the angle of contact between liquid and solid. As this angle decreases from 180 to 0 degrees, the wettability increases. Although the angle of wetting of water on highly purified cellulose is virtually zero degrees, the angle on wood is extremely variable and unpredictable owing to chance deposits of resinous materials and even occluded air on the cell walls. Even if the capillary walls were pure cellulose, the nonuniform cross section of the continuous capillary paths makes it difficult to calculate the degree of penetration. If a constriction occurs at some point in an otherwise uniform tube, penetration of the liquid beyond the constriction may not occur. The pit system in series with the fiber cavities is, in effect, a constriction; and hence if dry wood is immersed in a treating solution, capillary flow may not extend beyond the first pit system. Capillarity may occur in the cell walls; however, as mentioned previously, this may be looked upon as a type of solution rather than flow through a tube.

Pressure permeability occurs in response to an externally applied pressure. In completely saturated wood, flow can be considered to occur much in the manner of laminar flow in a tube; that is, Poiseuille's equation can be taken to represent the relationship between flow, pressure, properties of liquid, and size of capillary. The analysis becomes difficult if air-water interfaces are present, for, in addition to the frictional resistance offered by the capillaries, there is the additional resistance, due to capillarity, offered by these interfaces. In wood, the flow is not analogous to the flow in a single tube, but rather to the flow through a number of tubes in series-parallel combination. Stamm [45] has shown that the total flow depends on the applied pressure, viscosity, size and number of capillaries, and the type of arrangement of the capillaries—that is, whether they are in series, parallel, or in some combination thereof. Since laminar flow depends on the fourth power of the radius of the tube, it is apparent that the radius will be very important in determining the resistance to flow. In fact, Stamm has shown that for liquid water permeating wood, only about 0.02 per cent of the pressure flow is through the cell walls (where the effective capillary radius is of the order of Ångstroms), while the remainder is through the series and parallel combination of the microscopically visible capillaries, chiefly the cell-cavity pit system. The effective radius of the permanent pores in the pit membrane is much smaller than that of the cell cavity; therefore, it is these

pores that determine the permeability of the wood. The effective radius of the permanent pores is, however, very variable, especially in heartwood, because of the frequent complete or partial clogging up with resinous materials or because of the action of a valve-like membrane, the torus. Since the fourth power of an extremely variable and unpredictable quantity is involved, quantitative predictions of the pressure permeability of wood cannot be made. Only qualitative generalizations can be made,[45] as follows:

(1) Since permeability depends primarily on the effective radius of the pores in the pit membrane and insignificantly on the radius of the cell cavity, the permeability should be relatively independent of the density of the wood, which affects primarily the size of the cell cavity. This is especially true with the heartwood of different species where the size of the pores in the pit membrane is even smaller than that of the pores in sapwood, owing to partial clogging by resins.

(2) Since clogging of pit pores is less likely in sapwood, it is much more permeable than heartwood. Johnston and Maass [22] report that the rate of pressure permeation of water through Norway pine sapwood is 200 times that for heartwood.

(3) Owing to the greater number of pit membranes in the transverse direction than in the longitudinal direction per unit length of wood, the resistance to flow in the transverse direction is much greater than that in the longitudinal direction. For this reason, pressure treatment of wood with solutions is much more effective in the longitudinal direction.

(4) Because of the occurrence of tubes or vessels of relatively large radius in hardwoods, such woods would be expected to have a greater permeability than softwoods. Although some do, in other hardwoods the vessels are clogged with resinous materials called tyloses. Such woods have greatly reduced permeabilities.

Diffusion occurs in response to a difference in concentration or activity. The phenomenon of diffusion in wood is undoubtedly the most complex of the three types of flow that can occur in wood; and yet, paradoxically, it is the flow that is most amenable to analysis, because the factors involved are not as variable as those involved in the other types of flow.

Diffusion through a number of tubes depends only on the total cross-sectional area over which diffusion occurs and is independent of the size of the individual capillaries. In this respect, it differs from pressure permeability, which depends on the radius of the tubes through which permeation is occurring. Since the diffusion through capillaries of a given type depends only on the combined cross-sectional area, if there are enough of these in parallel, that is, if the fractional cross-sectional area of this type of capillary is large, then, even though this area is composed

of individual areas that may be extremely small, diffusion through this class of capillaries can be appreciable. Thus all capillary classes in wood, regardless of individual size, can contribute to the diffusion provided their combined cross-sectional areas are large; in other words, provided there are many of them connected in parallel. The network system is very complex, for it is composed of capillaries connected in series and parallel. Stamm [45] assumes that diffusion through the intricate network in wood is analogous to the conduction of an electric current through the same capillary structure. The diffusion through capillaries in parallel is the sum of the diffusion in each single capillary, and that through capillaries in series is the reciprocal of the sum of reciprocals of the diffusion through each capillary. The analysis which is confined only to the structurally simple softwoods and which neglects the small contribution by the medullary rays, assumes that diffusion occurs through cell cavities (tracheids) in series with the parallel combination: pit system-transient-cell-wall capillaries. A continuous passage through the transient-cell-wall capillaries is in parallel with the foregoing combined path. The pit system in turn is a complex system by itself. It consists of a pit chamber in series with the parallel combination of permanent and transient capillaries in the pit membrane. In wood of unit dimensions, there are present a definite number of capillaries of each type arranged in the above-described manner. To calculate the contribution of each capillary class, there are required the values for the total cross-sectional area and the average length of each class connected in parallel, the length, and number per unit length of wood, of those capillaries connected in series. These structural data have been obtained by Stamm,[38, 45] by use of classical physiochemical methods.

Consider first the diffusion of a water-soluble solute through water-saturated wood. Stamm's calculations show that the longitudinal diffusion decreases almost linearly with increasing density of wood. This results from the predominant influence of the void cross section of the cell cavities. This void cross section decreases with increasing density. For normal woods, longitudinal diffusion is about 15 times greater than the transverse diffusion. The calculated diffusion constants through wood are in good agreement with the experimentally determined values.[12, 13, 45] In Stamm's theoretical analysis, it is assumed that the medullary rays do not contribute to the overall diffusion. This is tantamount to assuming that the diffusion rates in both the tangential and radial directions are equal. Burr and Stamm,[11] however, find that the radial diffusion is generally slightly greater than the tangential diffusion (about 50 per cent). Furthermore, they find that the electrical conductance is greater than the tangential conductance by about the same amount. This latter observation confirms one of the basic assumptions in Stamm's calculations.

The analysis of the phenomenon of diffusion of moisture through wood is more difficult than that of the diffusion of salts through water-saturated wood. Whereas in the latter, the capillary dimensions remain constant, during the diffusion of water through wood, as during drying, some of the capillary dimensions change (Ref. 45, p. 12). The analysis is simplified by arbitrarily subdividing the block into a finite number of laminae in the direction of the moisture gradient, each of which will have its own capillary dimensions depending on the moisture content of the lamina. A further complication is the recognition that moisture diffuses not only as vapor in the cell cavities, but also as bound water in the cell walls. This dual nature of moisture diffusion has been recognized by several investigators.[30, 45] Rees and Buckman note that although the pressure permeability of gases through red oak is much greater than through bur oak, the rate of drying of red oak is less than that of bur oak. They attribute this to the greater proportion of the relatively slow bound-water diffusion in the denser red oak.

The vapor diffusion occurs in response to a vapor-pressure gradient, whereas the bound-water diffusion seems to occur in response to a moisture-content gradient.[45] The vapor diffusion is expressed in terms of the bulk diffusion constant of water in air. This constant is readily available in the literature. The choice of the bulk diffusion constant of the bound water is the least secure of the various assumptions in Stamm's analysis. It is assumed that the diffusion of such moisture is given by the Einstein equation for the diffusion of spheres out of contact with each other.[17] Owing to the dual nature of moisture diffusion in wood, the effect of density, temperature, and moisture content will depend on the effect of these variables on the vapor diffusion and bound-water diffusion. For densities less than 0.25 gram per cubic centimeter, the transverse moisture diffusion increases rapidly with moisture content. This is due to the fact that a greater proportion of the total diffusion in such woods is vapor diffusion. For normal woods, and when the moisture content of the wood is greater than 6 per cent, the diffusion constant is approximately independent of the moisture content. With increasing density, the diffusion constant decreases. This is because with increasing density, the proportion of the total diffusion due to bound-water diffusion increases, and this type of diffusion is slower than vapor diffusion. The overall diffusion constant of moisture through wood increases with increasing temperature. The rate of increase, however, is greater for woods of low density, for in such species vapor diffusion predominates.

The agreement between the values of the moisture-diffusion constant calculated from capillary-structure considerations and from laboratory-obtained drying data is surprisingly good, considering the many assumptions and approximations in the analysis. Since the analysis was made

only for softwoods, it would be expected that agreement for hardwoods would not be so good. Furthermore, since the analysis neglected the contribution due to the medullary rays, it would be expected that drying rates should be slightly greater in the radial direction than in the tangential direction. In confirmation of this, Rees and Buckman [30] found that the ratio of longitudinal to radial diffusion is about 5, whereas that for the longitudinal to tangential diffusion is about 9; that is, radial diffusion may be as much as 50 per cent greater than tangential diffusion.

With a knowledge of the diffusion constant of water in wood, the drying times under different conditions can be calculated by the methods described by Tuttle [61] and by Sherwood.[37] These calculations, of necessity, assume efficient air circulation. In actual large-scale drying operations, the circulation of the air through the lumber pile is undoubtedly not ideal, and therefore the calculation of the drying times gives the minimum drying periods.

Above the fiber saturation point, the cell walls are saturated and the aqueous relative vapor pressure of the cell cavities is approximately unity. On drying such a block of wood, a wet line develops that recedes toward the center of the block with time. On one side of the wet line, the moisture content would be expected to drop off rapidly towards the drying surface; on the other side, the moisture content would be expected to rise rapidly and then remain constant. In actual drying, however, there is no sharp break in the moisture content-distance curve. Hawley [10] has shown that the shape of this distribution curve depends on the degree with which the cell cavities are filled with water. If the cell cavities are completely filled, free water cannot flow. A sharp break occurs at the fiber saturation point. The moisture content rises very rapidly to a value corresponding to the moisture content of the original green wood. On the other hand, if air bubbles of appreciable size are present in the cell cavities, that is, if the cells are only partly filled, a break does not occur at the fiber saturation point; but, instead, the moisture content increases progressively and approaches the original moisture content. The overall diffusion, here, seems to involve the diffusion of free water. As free water leaves the surface cavities, air-water interfaces are created in the permanent pores of the pit membranes. The water is in this case under tension by an amount depending on the size of these pores. To reduce this tension, the air bubbles in the interior cavities expand and, in so doing, enable water to flow out of these cavities. Thus the flow of free water above the fiber saturation point is really governed by the diffusion of the moisture below the fiber saturation point.

In actual treating processes, combinations of these three types of flow generally occur. Consider the simple method of treating wood by soaking it in the treating solution. If the wood is initially air-dry, some flow by

capillarity occurs; however, the building up of back pressure may offer resistance to such flow. Diffusion of a solvent into wood frequently precedes the diffusion of a solute. It is thus often desirable when a water-soluble solute is to be deposited in wood by diffusion, to treat the wood in the green state. Stamm and Seborg[53] found that sheets of 1/16-inch Douglas-fir veneer containing 100 per cent of water on the basis of the dry weight of the veneer took up 25 per cent of the weight of the dry wood of phenolic resin in 8 hours from a 50 per cent aqueous solution. Similar veneer containing only 6 per cent moisture required 30 hours to take up the same amount of resin. Green veneer, however, is not always available. Impregnation times for air-dry veneer can be reduced by using high-pressure permeation. The resistance to flow resulting from back pressure can be reduced by evacuating the wood before immersing it in the treating solution. Air-dry wood is often impregnated with creosote by first evacuating the wood in a cylinder, and then by running in creosote under high pressure.

REFERENCES

1. Babbitt, J. D., *Can. J. Research*, **20**, 143 (1942).
2. ———, Private communication, 1948.
3. Berkley, E. E., *Text. Research J.*, **18**, 519 (1948).
4. ———, and Windyard, O. C., *Text. Research J.*, **18**, 519 (1948).
5. Barkas, W. W., *Proc. Phys. Soc.*, **48**, 1 (1936).
6. ———, *Proc. Phys. Soc.*, **48**, 576 (1936).
7. ———, *Trans. Faraday Soc.*, **35**, 1388 (1939).
8. ———, *Trans. Faraday Soc.*, **36**, 824 (1940).
9. ———, *Trans. Faraday Soc.*, **37**, 535 (1941).
10. Brunauer, S., "The Adsorption of Gases and Vapors," Princeton, N. J., Princeton Univ. Press, 1943; see also Emmett's paper in Vol. **VI**, this series.
11. Burr, H. K., and Stamm, A. J., *J. Phys. Chem.*, **51**, 240 (1947).
12. Cady, L. C., "Molecular Diffusion into Wood," Madison, Wis., Univ. of Wisconsin Library, 1934.
13. ———, and Williams, J. W., *J. Phys. Chem.*, **39**, 87 (1935).
14. Campbell, W. B., *Can. Forest Service Bull.*, **84** (1933).
15. Davidson, G. F., *J. Text. Inst.*, **18**, T175 (1927).
16. Eyring, H., and White, H. J., *Text. Research J.*, **17**, 523 (1947).
17. Einstein, A., *Ann. Physik*, **17**, 549 (1905).
18. Haines, R. S., and McIntoch, R., *J. Chem. Phys.*, **15**, 28 (1947).
19. Hawley, L. F., *U.S. Dept. Agr. Tech. Bull.*, **24B** (1931).
20. Hermans, P. H., "Contributions to the Physics of Cellulose Fibers," New York, Elsevier Publishing Co., Inc., 1946.
21. Huber, B., and Prütz, G., *Holz Roh-u. Werkstoff*, **1**, 377 (1938).
22. Johnston, N. W., and Maass, O., *Can. J. Research*, **3**, 140 (1930).
23. Kerr, T., and Bailey, I. W., *J. Arnold Arboretum*, **16**, 273 (1935).
24. MacInnes, A. S., West, E., McCarthy, J. L., and Hebbert, H., *J. Am. Chem. Soc.*, **62**, 2803 (1940).
25. McBain, J. W., "Sorption of Gases and Vapors," London, George Routledge and Sons, Ltd., 1932.
26. Meyer, H. H., "Natural and Synthetic High Polymers," New York, Interscience Publishers, Inc., 1942.

27. Meyer, H. H., and Misch, O., *Helv. Chim. Acta,* **20,** 232 (1937).
28. Meunier, A., and Guyot, O., *Compt. rend.,* **188,** 506 (1929).
29. Pauling, L., *J. Am. Chem. Soc.,* **67,** 555 (1945).
30. Rees, L. W., and Buckman, S. J., Univ. of Minnesota Agr. Expt. Sta., *Tech. Bull.,* **108** (1935).
31. Ritter, G. J., *Ind. Eng. Chem.,* **17,** 1194 (1925).
32. ———, *Paper Trade J.,* **101** (18), 93 (1935).
33. ———, *Paper Trade J.,* **108,** 33 (1939).
34. Ritter, G. S., in "Cellulose and Cellulose Derivatives," ed. E. Ott, New York, Interscience Publishers, Inc., 1943.
35. Soeman, J., *Ind. Eng. Chem.,* **37,** 43 (1945).
36. Schulz, E., *Z. physik. Chem.,* **B52,** 23 (1942).
37. Sherwood, T. K., *Ind. Eng. Chem.,* **21,** 12 (1929).
38. Stamm, A. J., *J. Agr. Research,* **38,** 23 (1929).
39. ———, *Ind. Eng. Chem. Anal. Ed.,* **1,** 94 (1929).
40. ———, *J. Am. Chem. Soc.,* **56,** 1195 (1934).
41. ———, *Physics,* **6,** 334 (1935).
42. ———, *Ind. Eng. Chem.,* **27,** 401 (1935).
43. ———, "Colloid Chemistry of Cellulosic Materials," *U.S. Dept. Agr. Misc. Pub.,* **240** (1936).
44. ———, *Ind. Eng. Chem.,* **30,** 1280 (1938).
45. ———, "Passage of Liquids, Vapors and Dissolved Materials Through Softwoods," *U.S. Dept. Agr. Tech. Bull.,* **929** (1946).
46. ———, and Hansen, L. A., *J. Phys. Chem.,* **41,** 1007 (1937).
47. ———, and ———, *J. Phys. Chem.,* **42,** 209 (1938).
48. ———, and Loughborough, W. K., *J. Phys. Chem.,* **39,** 121 (1934).
49. ———, and ———, *Trans. Am. Soc. Mech. Engrs.,* **63,** 329 (1942).
50. ———, and Millett, M. A., *J. Phys. Chem.,* **45,** 44 (1941).
51. ———, and Seborg, R. M., *Ind. Eng. Chem.,* **28,** 1164 (1936).
52. ———, and ———, *Ind. Eng. Chem.,* **27,** 1480 (1935).
53. ———, and ———, *Ind. Eng. Chem.,* **31,** 897 (1939).
54. ———, and ———, *Mech. Eng.,* **67,** 25 (1945).
55. ———, and Tarkow, H., *J. Phys. Coll. Chem.,* **51,** 493 (1947).
56. ———, and Tarkow, H., Paper presented before Colloid Symposium, Cambridge, Mass., June 23, 1948.
57. Tarkow, H., Unpublished work.
58. Tiemann, H. D., "Wood Technology," New York, Pitman Publishing Corp., 1942.
59. ———, Research Rep., Forest Service, Madison, Wis., 1944.
60. Trogus, C., and Hess, K., *Z. physik. Chem.,* **15B,** 157 (1932).
61. Tuttle, F., *J. Franklin Inst.,* **200,** 609 (1925).
62. Weatherwax, R. C., and Stamm, A. J., *Trans. Am. Soc. Mech. Engrs.,* **69,** 421 (1947).

CELLOPHANE

Charles M. Rosser

American Viscose Corp., Fredericksburg, Va.

Early History

The word *cellophane* has become, through common usage, a term to denote the films or foils of cellulose hydrate regenerated from cellulose solutions such as viscose and cuprammonium. A detailed description of the early developments and history of cellophane are set forth in a German book by Halama,[1] which was published in 1932. Although Brandenberger is rightfully considered the father of the cellophane industry, the general idea of forming films from Cross and Bevan's[2] viscoid (viscose) was disclosed in a British patent by Stearn[3] in 1898. Stearn proposed that the viscose be forced through a fine slit into an ammonium chloride solution for regeneration. In 1902, British Patent 2,529 mentioned films made from cellulose.

The first practical production of regenerated cellulose film[4] was made by Brandenberger in 1908, while employed in a textile works in Thaon-les-Vosges, France. His original thought was to use his product in the photographic industry. It was a film about ten times thicker than present films. Brandenberger coined the word cellophane to describe this material: *cello* from cellulose and *phane* from diaphane (diaphanous), for its clarity and transparency; *cellophane* = transparent cellulose. He had a process and a product but little market. By 1912 the film thickness had been lowered to 0.02 mm (0.0008 in), and a use in packaging was developed for the product. The cellophane industry was established, as S.A. La Cellophane in Paris, in 1913. This early date may interest those who consider cellophane a modern development.

La Cellophane licensed duPont de Nemours, in 1923, to use its process in the United States. In 1926, the firm Kalle and Company was licensed in Germany. Competition had sprung up in Belgium by Sidac (La Société Industrielle de la Cellulose) and in Germany by Wolff and Company. Production was underway in England, Poland, Holland, and Canada.

Before and during World War II, Japan was a producer. Cellophane is now also made in Sweden, Spain, Italy, Egypt, Czechoslovakia and Brazil. Current production in the United States alone is reported to be about 200 million pounds annually, serving some 5,000 uses.

Manufacture of Cellophane

In the first steps of cellophane manufacture by the conventional viscose process, the procedure is similar to the production of viscose bright rayon as outlined by Moss.[5] Briefly it is as follows: A high grade of cellulose pulp sheet (high alpha-cellulose content) is steeped in 18 per cent sodium hydroxide under controlled conditions. After pressing to a predetermined weight, the alkali cellulose sheets are shredded and aged at closely controlled temperature. The aged "crumbs" are reacted with carbon disulfide (xanthation), and the product is dissolved in dilute sodium hydroxide solution to form viscose. The usual filtration steps are observed, and after proper ripening, the viscose is ready for regeneration into cellulose hydrate.

From this point on, cellophane differs in method but not principle from rayon manufacture. In fact, the more recent continuous spinning of rayon resembles cellophane spinning (casting) more closely than it does the pot spun-rayon method. In the cellophane and continuous rayon methods, the foil or filament is a continuous entity from the extrusion point to the finished windup. Whereas the viscose rayon is extruded through fine holes into the coagulation bath, cellophane is cast through a long slit (about eighty inches in length).

In the Brandenberger process the viscose was cast free from the immersed lips of the hopper. The film sheet traveled unsupported for a short interval before it reached a submerged roller in the coagulating tank. In the Transparit process, developed by Wolff and Company and put on the market in 1923 in Germany, a competitive type of casting system was used. It was an adaption of the known method by casting on a moving wheel above the coagulating bath level. The layer of viscose was coagulated as the wheel submerged in the acid bath, and the formed film was stripped from the highly polished surface near the point of emergence from the liquid. The wheel was approximately eight feet in diameter. Otherwise the two processes closely resembled each other, except for slight differences in drying procedures. Of the two methods, it is believed the "free" casting is in general use today. However, from observations on some physical properties of a sample of Japanese film captured during the war, it appeared probable that the film had been cast upon a support.

In the conventional process the coagulating acid bath is followed by

more dilute acid regeneration, desulfurization, bleaching, and plasticizing baths. Between each set of processing baths are water washes to flush the successive chemical treatments from the film. Glycerin, in preference to other softeners, is most widely used, as it is odorless, tasteless, colorless, nontoxic, nonvolatile, hygroscopic, and relatively inexpensive. It is possible, by choice of other softeners, to produce a film which will retard the propagation of an applied flame.

For colored grades the dyeing is done on the casting machine, although colorless dried film may be dyed later on a special machine. The most frequently used colors are pink, red, amber, blue, green, and black.
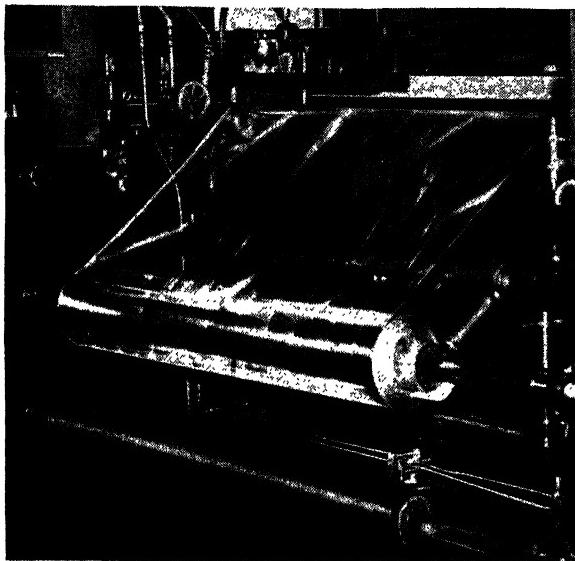
Figure 1. Cellophane "mill roll." The sheet is wound into rolls of suitable length.

Upon exit from the plasticizer bath, the film is pressed, usually by rollers, to remove excess liquid, and it is then passed to the dryer section. When one considers that the film at this point contains 70 to 80 per cent liquid, the drying requirements are important. From this step on the cellophane process conforms more to paper practice than to rayon. The conventional dryer consists of a series of staggered iron rolls, internally heated, over which the film passes as the water is removed. There is usually some provision made to carry off the moisture-laden air from the dryer area. It is essential that some control be kept over the film during this stage to produce a smooth clear sheet. The water in the cellophane is reduced to an optimum figure, and the sheet is then wound into rolls

of suitable weight (Figure 1). The water content is a minor percentage (six to eight per cent), but a very important factor in the usefulness of the film.

In the manufacture of the base film by the conventional method, a "grain" is cast and dried into the film structure in the direction in which the film is extruded and passes through the processing baths and over the dryers. There is little restriction in the transverse direction during manufacture. Whatever irregular grain is formed, is reflected in the later life and usage of the film. These effects may be observed by viewing cellophane in polarized light. As a result of the above handling during manufacture, cellophane possesses certain inherent properties. Its machine-direction tensile strength is greater than that in the transverse direction, under standard testing conditions; conversely, its machine-direction elongation is less than that across grain (transverse). Its resistance to tearing is greater across than with the grain. Changes in moisture content cause dimensional changes in the sheet, which are more pronounced in the transverse direction.

The major portion of cellophane film is coated with lacquers deposited from volatile solvents. The lacquered surface is hydrophobic and retards the passage of moisture vapor through the foil. The most common lacquer base (film former) is low-viscosity nitrocellulose, although many other bases have been considered. The usual solvents are the lower alkyl acetates, hydrocarbon diluents, and perhaps an added alcohol. Most compositions include a resin component, a wax, and a plasticizer for the film former. By variation in the solids composition, various desirable properties may be imparted to the finished product, such as heat seal (self-sealing with heat), water-vapor resistance, ink reception in printing, and resistance to immersion in water. The base film is quite impermeable to dry gases and resistant to passage of oils. This property was utilized in the last war in the production of mustard gasproof protective covers for military personnel.

To conduct the process in an economical manner, recovery of chemicals is necessary wherever possible. As in the rayon industry, the sodium sulfate is recovered and sold. Acid is also reclaimed and reused. The volatile solvents in the coating operation are recovered by adsorption on activated carbon, then distilled, and reused.

Grades of Cellophane

Cellophane is produced in the following weights or gauges: 300, 450, 600, and 800. This terminology is a carry-over from the European practice, in which the weights are listed in grams per square meter. The type 300 weighed 30 grams per square meter, the 450 was 45 grams, and so on.

The public encounters the 300-grade on cigarets, bread, etc. The 300-grade is produced in largest quantity since it gives the greatest coverage per unit cost. Where more body or rigidity is required, the 450-grade is used. The 600-grade is familiar to all as the backing on pressure-sensitive tapes. Some use is made of the 800-grade in tear strips on chewing gum and cigarette packages.

TABLE 1. COMPARISON OF MOISTUREPROOF AND
NON-MOISTUREPROOF CELLOPHANE GRADES

Non-moistureproof			Moistureproof		
Gauge	Approx. Thickness (in)	Approx. sq inches per lb	Gauge	Approx. Thickness (in)	Approx. sq inches per lb
300	0.0009	21,500	300	0.0010	19,500
450	0.0013	15,000	450	0.0014	14,000
600	0.0016	12,400	600	0.0017	11,600

Cellophane is divided into two general classes. One is the plain or so-called non-moistureproof grade (Table 1). The other is the moistureproof class. A further definition here is helpful. The non-moistureproof type is hydrophilic and has surfaces wettable with water; in addition, it is not resistant to the passage of moisture vapor. The moistureproof grade has hydrophobic coatings applied over the hydrophilic base film, rendering the surfaces water-repellent. The resistance of these coated films to the passage of moisture vapor may be varied from high to low so as to suit the particular need of the user. However, water-repellent films are not considered waterproof (or liquidproof). Where usage demands continued exposure to high moisture contents or immersion in water, an anchoring treatment is necessary before application of the usual coating, although it is possible to prepare a vaporproof, waterproof coating in one step.

For special uses, the product may be a combination of both classes. That is, the lacquer coating is applied to one surface only, so that the film has one water-wettable surface and one water-repellent surface. It may be produced in varying degrees of water-vapor resistance. At present, this type is being used for wrapping red meats in self service markets.

An untreated, plain or non-moistureproof stock would tend to stick together or block, thus making separation tedious and difficult. This is usually prevented by either treating the finished spun film with an anti-blocking agent deposited from organic solvents, or by the application of a sizing bath while the film is on the spinning machine. In either case, the hydrophilic character of the surface is retained.

The wet gel film, obtained before passage over the drying rolls, finds application in dialysis. At present it is popularly used in this country as

an osmotic membrane for the determination of molecular weights of polymers. The undried gel is shipped in water.

Specific Reports and Patents on Cellophane

As previously mentioned, the book by Halama gives a very complete picture of the early history of cellophane in Germany. Similarly, the most detailed reports of modern practice appear to be the studies made by the Allied teams in Germany, following World War II, e.g., the British B.I.O.S. Report No. 858 and the American F.I.A.T. Reports Nos. 956, 934, 553, 480 and 59.[6-11]

American Report. Field Information Agency Technical Report No. 956,[7] contains seventeen drawings and diagrams setting forth details of the casting machine, hoppers, dryers, and coaters. In addition, there are ten photographs of various phases of operations, including casing manufacture. As in other reports, compositions of the various processing baths are given relative to the speed of casting. Several lacquer formulas are listed, as well as processes for making anchored film, as described in the following patents:

German Patent 752,714 [See U.S.A. Application 339,493 (June 8, 1940)]. Anchoring of lacquer layers on hydrophilic foils. According to claim, "Ketendimeren" is intended as interlayer.

German Patent 696,168 (August 24, 1937). Process for manufacture of cellulose hydrate foils coated with waterproof layers. The process claims a thin interlayer of mixed polymers of maleic anhydride and vinyl derivatives in organic solvents, prior to lacquering proper.

German Patent 747,627 (August 5, 1942). Cellulose hydrate foils are treated with isocyanate solutions and then lacquered, or the isocyanates are added to the lacquers used for coating the foils.

German Patent 752,808 (January 15, 1943). Increase of the adhesive strength of moistureproof layers on cellulose hydrate foils. Prior to applying the moistureproof lacquer, the unlacquered foils are treated with thin solutions of N-N-alkylene ureas in organic solvents.

German Patent 753,191 (May 24, 1943). Process for increasing the adhesive strength of moistureproof layers of cellulose-hydrate foils. Ethylene-imine polymers are applied in aqueous solution to the cellulose-hydrate foil prior to drying; the foil is then lacquered in the usual manner.

The formula for moistureproof heat-sealing lacquer is listed in the report as follows:

MST Type

77.0 kg nitrocellulose E 730 spirit moist (50 kg dry)
0.5 kg KM resin special (I.G. rosin-maleic resi)
0.5 kg FLS resin (composition unknown)

    3.0 k Alkydal RD25 (I.G. alkyd resin)
  12.5 kg Uresin B (I.G. urea formaldehyde resin)
  22.3 kg dibutyl phthalate
  16.5 kg dicyclohexyl phthalate
   4.0 kg paraffin, m.p. 54°
 560.0 l solvent mixture (benzene and n-propyl acetate
        in weight proportion of 2 to 1)

Included in the report are also some laboratory testing and evaluation methods for the films.

British Report. Although lacking in illustrations and drawings, the British report, British Intelligence Objective Sub-Committee Report No. 858, is very complete in detail. Processes used by the main competitors in the cellulose-hydrate foil business, Kalle and Company and Wolff and Company, are well covered, as well as the cuprammonium process of J. P. Bemberg.

Viscose Process. Regarding the two different casting methods used by the German companies in the early days, it is worth noting that Kalle was still using the immersed-lip type exclusively, while Wolff was using both the immersed-lip casting and the drum method. It is reported that films manufactured by the two processes showed no differences in tensile strength and elongation, although the drum-cast film exhibited more slip. Hence the latter was diverted into plain film, and the former into coated film.

There is a short summary of the method and difficulties encountered by the Germans in the development of continuous steeping and pressing of the cellulose pulp in the caustic-soda system. Weighed quantities of pulp were agitated with a measured quantity of steeping soda until a slurry was formed. The slurry was fed at a constant rate to a screw-type press, whose speed was balanced to the inlet slurry to give a cake of desired composition. The cake was shredded, aged, and reverted back to the batch process in the usual xanthation step. Some continuous aging was also tried. In general, progress is represented by improvement and refinement of older methods of viscose cellophane manufacture.

Wolff and Company data on setting and regeneration baths are given as follows:

Immersed-Lip Type

Film Weight (gm/m²)	Speed (m/min)	Setting Bath H₂SO₄ (Vol %)	Setting Bath Na₂SO₄ (Vol %)	Temp. (°C)	Regen. Bath H₂SO₄ (Vol %)	Wash and Regen. Bath H₂SO₄ (Vol %)
30	32.0	15–16	16–17	42.0	7–8	4.5
40	21.5	16	16–17	42.0	8–9	5.0
60	15.3	16	16–17	45.0	8–9	5.0
120		[Could not be cast on this machine]				

Drum Type

30	22.2	8–8.5	12–14	30	6–7	3–5
40	15.3	8–10	12–15	32	7–8	4–5
60	8.3	10–12	14–16	32	7–8	4–5
120*	4.2	10–11	14–16	32	7–8	4.5

* Possible only with special high-caustic, high-cellulose viscose.

The British report discusses the finishing equipment in use in Germany. This includes the slitting and reeling machines and converting machines, such as bag formers and printers—both aniline and rotogravure. It would appear, that present practice in this country is far advanced in speeds and technique, such as described in the publication, *Modern Packaging Catalog.*

Other Processes of Manufacture. Up to this point, this review has been confined to cellophane produced by the viscose process. Regenerated cellulose was first manufactured by Bemberg, A.-G. in 1932. Their manufacture of the film by the cuprammonium process on a large scale began in 1936. No moistureproof film had been made up to 1946. The product was sold under the trade mark "Cuprophan." It was produced in lighter-weight grades than viscose film, i.e., 15, 23 and 34 grams per square meter. The product was used for wound dressings (from the 15-gram stock)—slit and woven, and overlaid with lint. It was also woven into belts impregnated with an unknown substance. Other outlets for "Cuprophan" were in cable insulation and ladies hats. For films of equal thickness, the viscose film was reported cheaper.

Technically, cellulose (wood pulp, cotton linters, rayon waste) was dissolved (9 per cent) in cuprammonium solution prepared from copper oxychloride, ammonia, caustic soda, and sodium arabonate solution. In contrast to the viscose spinning solution, the cellulose cuprammonia solution could be stored for months without substantial deterioration. The spinning solution was spun down through a slot in a hopper, through an air gap, into an aqueous caustic-soda setting bath, to form a gel film, which after washing was regenerated in a dilute sulfuric acid bath. After further washing with water, the film was plasticized with glycerin or triglycol and dried in the usual manner. Only half the plasticizer content was found necessary for this film, as compared with viscose film. Also of interest was the use of ion-exchange resins for recovery of the copper from solution.

In a field closely allied to cellophane, tons of seamless regenerated cellulose tubes are sold as casings for meats, cheeses, etc., or cut in narrow bands to be dried on as bottle closures. A major portion of these products are produced by the viscose process, differing mainly in the choice of an annular in place of a straight slot for extrusion.

Some tubing is manufactured by the denitration of cellulose nitrate to

form cellulose hydrate. In this process, nitrocellulose (cellulose nitrate) is dissolved in ether-alcohol mixture and extruded into tubular form. After washing, the tubing is denitrated (ester-splitting) by alkaline media (usually sodium hydrosulfide solutions). This process produces an excellent film and one difficult to match by the viscose process.

Regenerated cellulose films may also be produced by the deesterfication of other cellulose esters, i.e., the acetate. A recent patent [12] refers to this process as one for the production of regenerated cellulose sheet materials. The method comprises swelling the dry, nonfibrous sheet with an aqueous solution of acetone and saponifying the swollen sheet. The resulting regenerated cellulose sheet is claimed to have improved and substantially uniform directional strength and elongation characteristics.

REFERENCES

1. Halama, M., "Transparentfolien," Berlin, Bodenbender, 1932.
2. Cross, C. F., and Bevan, E. J., Brit. Patents 8,700 (1892) and 4,713 (1896).
3. Stearn, C. H., Brit. Patent 1,020 (1898); Germ. Patent 108,511, Class 29 (1898).
4. Brandenberger, J. E., Germ. Patent 230,558, Class 57b (1908).
5. Moss, S. A., Jr., in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. VI, p. 945, New York, Reinhold Publishing Corp., 1946.
6. Monbiot, M. F., et al., "Investigation of the German Regenerated Cellulose Film Industry," B.I.O.S. Rept. 858.
7. Kohorn, O., and Costa, J. L., "Cellulose Sheeting and Sausage Casing Machinery and Processes," F.I.A.T. Rept. 956.
8. Nash, R. W., "Moisture-proof Cellophane Equipment and Coating Formulae," F.I.A.T. Rept. 934.
9. Wade, W., "Cellophane and Sausage Casing Made at Kalle & Co., A.-G.," F.I.A.T. Rept. 553.
10. Leach, L., et al., "German Cellophane," F.I.A.T. Rept. 480.
11. Hoff, G. P., et al., "Regenerated Cellulose Fibers at I. G. Farbenindustrie A.-G. (Kalle)," F.I.A.T. Rept. 59.
12. Dreyfus, C., Andersen, B., and Kojevnilkoff, N. (to Celanese Corp. of America), U. S. Patent 2,459,927 (Jan. 25, 1949).

PROTECTIVE WRAPPINGS—
BRITISH APPROACH

C. R. Oswin

British Cellophane Ltd., Bridgwater, Somerset, England

THE FUNCTION OF transparent wrappings has passed almost imperceptibly from decoration to protection—protection of the wrapped product during its journey from the packer to the consumer. The transparent pack is a basic unit of "super-markets" and self-service stores, but the relatively costly wrappings would probably not be used for common wares if they did not reduce loss by deterioration and waste.

Much of the material of ordinary merchandise is perishable; that is, on keeping it will deteriorate. The agent of deterioration is sometimes air, for example in the oxidation of fats; but damage is more often caused by air-borne matter such as water vapor or mold spores. Some wares may, of course, be particularly susceptible to mechanical damage, but mere wrapping affords no protection against such harm.

The commonest single cause of deterioration is a change in water content. Many commodities are hygroscopic, and change their water content according to the relative humidity of the ambient atmosphere, losing, at certain water contents, their palatability or texture. The transparent wrappings which are used commercially do not entirely prevent this change, because they are not perfect barriers against water vapor; but some of them offer such high resistance to the passage of water vapor that the rate of change of water content becomes very slow. The change is thereby held within an acceptable amount for as long as the package is likely to be kept before use.

The rate of change of water content in wrapped goods can be estimated [1] to see whether a given design of package will be adequate for its intended commercial life. Several factors combine to determine the life of the package, and they may be considered as functions of (1) the material to be packed, (2) the wrapping material used, and (3) the climatic conditions to which the packages are subjected. It is possible to test the efficacy of sample packs merely by exposing them to typical

atmospheric conditions,[2] but this method is slow and extravagant in comparison with a theoretical analysis of the package efficiency, backed by one or two accelerated storage tests. The basis of the theoretical examination is the analogy between a package (consisting of a water-vapor-resistant barrier surrounding a material of known water capacity, the whole being exposed to a water-vapor pressure gradient) and an electrical circuit, consisting of a condenser of known capacity being charged by a steady potential applied through a resistance. It is already known that the rate of charge of such a condenser would be exponential, and the exponential rate of decay was introduced independently into three package-life studies in 1945.[3]

The Nature of the Material to Be Packaged

Some materials are too sensitive to changes in water content to be packed in flexible wrappings at all. They are usually crystalline materials which undergo a change of form when they gain or lose small quantities of water, and the changes occur abruptly at definite relative humidities. The complete change of crystalline form is accompanied by a relatively large change in the amount of water present, but deterioration of the packed material becomes evident when only a small proportion of the change has taken place. For example, pure sucrose does not absorb water from the atmosphere at all when the relative humidity is below 85 per cent,[4] but in contact with atmospheres of slightly higher humidity, absorption is rapid and the deliquescence produces caking of the crystals. Similar behavior is shown by sodium chloride, which is hygroscopic above 75 per cent relative humidity. The behavior of typical crystalline materials under different conditions of relative humidity is indicated in Figure 1.[5]

When the hygroscopic material has a colloidal nature, e.g., an insoluble gel, the effect of changing relative humidity is always more gradual. At no stage is there a sudden jump in the equilibrium water content for a slight increase in relative humidity, but the increase is gradual, and the graph of water content against relative humidity (Figure 2) is a characteristic ogee, the so-called "sigmoid isotherm."[6] The adsorption and desorption isotherms often diverge slightly, showing a "hysteresis" loop, but it is not necessary to allow for this in simple package-life calculations. The values of the equilibrium water contents may also vary for different samples of the same gel, according to the history of the material, but these variations can be ignored in simple package-life equations.

The change in water content of the colloidal materials may be accompanied by a change in dimensions or in physical properties, but

the change is gradual. The swelling or contraction is not catastrophic unless there is a sudden collapse of the structure, accompanied by syneresis or "crazing." When deterioration of the colloid material is caused by change in water content, it is generally spread over a range of humidity. The packed commodity may spoil eventually by loss of moisture, by mold growth, or other means; but always a certain amount of change in water content can be tolerated [7] before deterioration is serious, and it is this characteristic behavior of colloid materials which makes the use of protective wrappings profitable. Not only is the change tolerated, but it produces a slowing down in the rate of penetration. The water-vapor-resistant wrapping so retards the change of
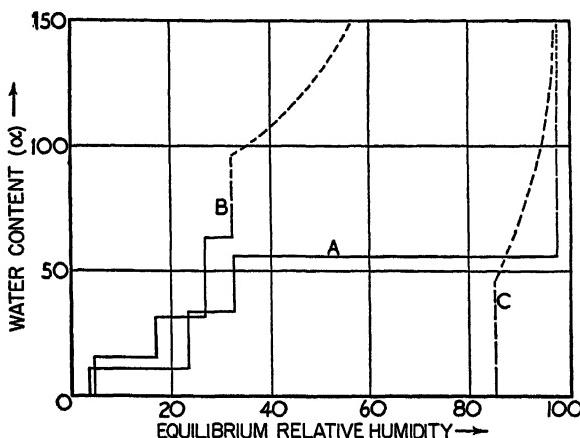
Figure 1. The water relations of some crystalline materials: (A) copper sulfate, (B) calcium chloride, (C) sucrose.

water content in the wrapped material that deterioration is avoided for a period of time sufficient to cover the journey from the packer to the user. When the perishable material is not hygroscopic but reacts, like iron, with moist air, deterioration may be retarded by placing an absorbent material such as silica gel inside the water-vapor barrier.

Naturally, the greater the difference between the water content of the goods at the time of packing and that at which deterioration becomes serious, the longer the package can be expected to last under given conditions and this factor, the difference between the moisture content at packing, a_0, and that at which deterioration appears, a, has been called the "permissible water change." The dry weight of material, W, enclosed in a package of given size will also affect the life of the package, since the rate of ingress of water vapor is not affected by the weight of filling amongst which it is distributed. In practice this

does not mean that it is necessary or desirable to compress a pound of material into a half-pound package, but that with a given area of wrapping, a, it is desirable to enclose the maximum weight of goods. This calls for avoidance of packages which depart widely from the shape of a sphere or cube. In order to simplify the calculation, all water contents are expressed as percentages of the dry weight of the packed material.

In developing the package-life equation, it is necessary to transform the humidity gradient into a function of the water content, and as will be seen from Figure 2, the relation is complicated. For normal commercial purposes, however, no serious error is introduced by assuming
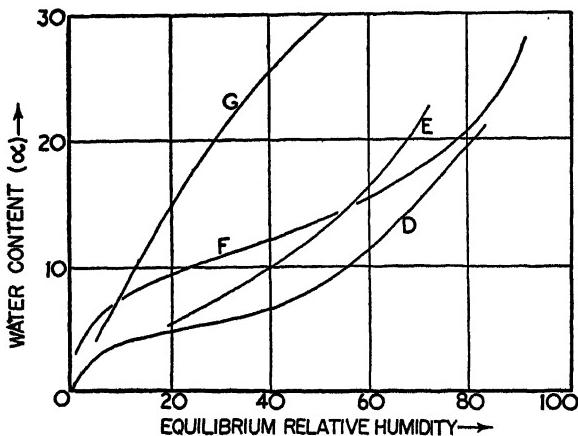
Figure 2. The water relations of some colloidal materials—adsorption isotherms only: (D) breakfast cereal, (E) tobacco, (F) starch, (G) silica gel.

that the isotherm is a straight line. The slope of such a hypothetical linear isotherm, μ, is the average increase in equilibrium water content corresponding to a 1 per cent increase in relative humidity of the atmosphere, and it expresses the "avidity" of the goods for water.

The Nature of the Wrapping Material

The effectiveness of the package in preserving its contents against changes in water content also depends on the resistance offered by the wrapping to the passage of water vapor. The resistances of the different commercial wrappings which are available vary over several orders of magnitude, and it is not always necessary to use the most resistant wrapping.[8] Laboratory measurements of the resistance of the wrappings to water vapor are usually made by sealing a sheet of the wrapping over the mouth of a cup which contains water, an aqueous solution, or a

desiccant. More or less elaborate means are adopted to seal the edge of the sheet according to the frequency with which tests are made.[9] The closed cup is weighed, and placed in an atmosphere which is either drier or wetter than the inside of the cup; after allowing a suitable induction time, the cup is weighed periodically to find the rate of loss or gain of water. For the test to be useful in manufacturing process control, it must be completed in 24 hours, and the more resistant wrappings do not then give sufficient change of weight for accurate determination unless a temperature of about 40°C is maintained. Tests at 0°C and −20°C, although desirable, are not yet practicable because of the slow loss.

There are many variants of the test, but the difference in results is generally of a smaller order than that between different grades of wrapping.

Many different units are used[10] for expressing the resistance (or its reciprocal, the permeability) of a wrapping to water vapor. Most of them are weight changes under specified conditions, but for package-life studies it is better to have a convertible unit, and the resistance of wrappings to vapor has been defined[8] in cgs units as the pressure in dynes/sq cm required to force 1 cc of water vapor at a pressure of 1 dyne/sq cm through 1 sq cm of the wrapping in one second. The resistance in these units, divided into 69,000, gives the GFMVTR in g/100 sq in/day at 40°C, at 5 to 91 per cent relative humidity. If the resistance of the wrapping is so high that it can only be measured accurately at temperatures approaching 40°C, it can be corrected back to more normal temperatures, since the relation between resistance and absolute temperature is an exponential one. The figures quoted below are corrected to 25°C.

Papers, glassines and other fibrous sheets offer little resistance to the passage of water vapor unless they are waxed or coated. *Waxed-papers* and *-glassines* have a fairly high resistance to water-vapor (12,000 cgs units) but they are easily damaged by creasing or crumpling. *Varnished papers*[11] or papers coated with a plastic such as polythene do not suffer from this disadvantage, and they may show resistances up to 5000. *Regenerated cellulose*[12] film is made in thicknesses from 0.0008 to 0.0016 in by dissolving xanthated cellulose in dilute caustic soda, extruding the solution so formed through a linear orifice and regenerating the cellulose by means of buffered sulfuric acid solution. The film is impervious to oils and greases, but it is scarcely more resistant to water vapor than is paper. The application of a nitrocellulose solution containing plasticizers, blending agents and waxes, confers on the so-called "moistureproof" cellulose film a high resistance to water vapor (up to 45,000), thus combining protection with the desirable mechanical properties of cellulose sheet. Intermediately moistureproof films for certain

uses may have a resistance of about 4000. *Cellulose acetate* [13] films, made by spreading solutions of plasticized cellulose acetate from organic solvents on to a polished support and evaporating the solvent, show a low resistance (250) to the passage of water vapor. Acetate films made by extrusion, and films of acetate-butyrate have a similar resistance. Normal thicknesses are 0.001 to 0.003 in.

The thermoplastic films, unlike coated papers and cellulose films, are resistant to water vapor throughout their thickness. They are flexible, but because they are always plasticized so as to seal by heat, they show more stretch on printing and bag-making machinery than do the cellulosic wrappings. They are more resistant than coated cellulose to damage by creasing or scratching, but tend to be less stable. *Rubber hydrochloride* [14] film is made by casting a plasticized solution in hydrocarbons on to a support and evaporating the solvents. Freshly made film about 0.0009-in thick has a high resistance to water vapor (20,000) but on exposure to air or light the film slowly deteriorates, becoming brittle, and losing more than half its resistance to water vapor. *Polyethylene* [15] films deteriorate much more slowly but usually have a lower initial resistance to water vapor (6000 at a thickness of 0.001 in). They are made by extruding molten polymer, or by calendering. Thicker films, up to 0.003 in, may show resistances comparable with coated regenerated cellulose films. They are not perfectly transparent. Flexible films made from *vinyl* [16] polymers show only moderate water vapor resistance, ranging from 1000 for plasticized polyvinyl chloride to 200 for polyvinyl alcohol. *Polyvinylidene chloride* [17] films are also made by band casting; they show exceptionally high resistance to water vapor (120,000). As yet, however, they are only produced on a small scale. Wrappings of higher resistance than those listed can be made by *laminating* [18] two or more plies; any of the above wrappings may be laminated together, to other wrappings, or to *metallic foil*. The increased thickness of laminated materials makes them more difficult to use on automatic wrapping machines, but they have the advantage of being applied in one operation.

Alternatively, where the resistance afforded by a single wrapping proves inadequate, high values can be obtained by using several resistances in series. These can take the form of multiple wraps, but more economical results are achieved with bulk overwraps. The penetrable area of the wrapper, a is also a factor in the package-life equation, and where wrappings of different resistances are used in different parts of the pack, the resistances can be summed in parallel by the usual formula.[19] This is necessary with window packs and with packs of waxed paper or regenerated cellulose film, where part of the wrapping has been damaged by scratching or by heat-sealing.[20] The sum of areas and resistances is expressed by the term $\Sigma a/r$ in the package-life equation.

The Effect of Storage Climate

Two components of the storage climate affect package life. One is temperature, the other relative humidity. The relative humidity determines the equilibrium moisture content (a_s) which the contents of the package approach,[6] while the temperature affects the rate at which the approach is made.[3] In practice, the storage humidity is not constant, but Lavers[21] has shown that the humidity inside a package oscillates about a curve, which is roughly exponential when the external fluctuations of humidity are regular. It is sufficient for commercial purposes to take the average temperature and humidity for climates to which the pack is likely to be subjected.

Attempts have been made to systematize climatic types, notably by Köppen and Thornthwaite,[22] but seasonal variations complicate the classification. Brooks[23] has calculated an index of climatic deterioration, but his index only applies to deterioration of materials under the influence of a high humidity, and does not take into account the fact that many commodities, for example cigarettes or bread, spoil by losing some of their water to the air. In general, meteorological data are needed for each country to which packs are likely to be sent. It must be borne in mind, however, that packs are normally stored indoors, and that if the average outdoor temperature falls below 65°F, artificial heating may raise the indoor temperature and reduce the storage humidity.

The effect of humidity of the storage atmosphere is allowed for in the package life equation through the term, a_s, the moisture content which the goods would attain if they were exposed unwrapped to the storage atmosphere. The total change which can take place, $a_s - a_0$, is called the "potential water change." The ratio of (potential water change $minus$ actual water change at any time) to (potential water change) is called[24] the "dishumor", ð ,* since it denotes the relative amount by which the contents of the package are still out of humor with the storage atmosphere at any time after exposure has begun.

Temperature affects package life in two ways—firstly by increasing the vapor pressure of water which corresponds to a given relative humidity, and secondly by reducing the resistance of the wrapper. The variation of water vapor pressure (p) with temperature is tabulated, and it has been found[3] that the resistance of "moistureproof" cellulose film is inversely proportional to $p^{1.6}$. The package life is thus inversely proportional to $p^{2.6}$. This sensitivity to temperature is reflected in the fact that a decrease of 8°F in the storage temperature doubles the shelf life.

* This is the Anglo-Saxon *thorn*, pronounced like the letters *th*.—*Ed.*

COLLOID CHEMISTRY

The Package-Life Equation

The full equation which has been developed for estimating the life of a package, at least up to the half-life, is:—

$$t = \frac{W\mu p_m^{1\cdot 6}}{p_s^{2\cdot 6}\Sigma a/r} \log_e \frac{(\alpha_s - \alpha_0)}{(\alpha_s - \alpha)}$$

but for rapid calculation this can be simplified without introducing serious errors. By taking weighted averages for a and r and correcting r to a standard temperature, a number of factors can be combined together in a single factor, K, which expresses the inherent absorbence of a package. The temperature factor, $p_s^{2\cdot 6}$, runs closely proportional to the fourth power of the temperature in degrees Centigrade over the range 15 to 35°C (which is the normal range of indoor storage temperature). Further, there is no need to allow for extensions of package life beyond about half-way to the equilibrium moisture content, α_s, since if the contents of the package did not spoil by that point there would be little need for protection by wrappings. So the term $\log_e (\alpha_s - \alpha_0)/(\alpha_s - \alpha)$ can be replaced by the approximation $2(1-\delta)/(1+\delta)$. Then, collecting all constants into the efficiency term K, the equation becomes:

$$t = \frac{Kr\,(1 - \delta)}{T^4\,(1 + \delta)}$$

Maximum values of r have already been quoted: in practice they will be somewhat reduced, especially if there is abrasion, creasing or heating of the wrapping during application. Some typical values of K are quoted in the table below:

TABLE 1. THE FACTOR K FOR
SOME TYPICAL PACKAGES

Biscuits ($\frac{1}{2}$ lb)	32,000
Cigarettes (20)	5,400
Dried egg (5 oz)	16,000
Breakfast cereals (12 oz)	24,000
Rice ($\frac{1}{2}$ lb)	13,000
Tea ($\frac{1}{4}$ lb)	8,200
Cornflour (1 lb)	21,000
Starch ($\frac{1}{2}$ lb)	15,000

This enables the life of packages in different wrappers to be assessed. The graph in Figure 3 shows how the equation was followed in practice by packets of twenty cigarettes stored at 32 per cent relative humidity and 95°F.

As an additional safeguard, laboratory tests should always be made by exposing sample packs to known, fixed conditions.[25] The equation does not apply to crystalline materials, to materials which crystallize or grain, or to materials which themselves offer a high resistance to the pas-

sage of water. Problems of this nature arise in the wrapping of sugar boilings, or of moist foodstuffs which would grow mold unless an inefficient wrapper allowed the surface "crust" to become too dry to support mold growth.

Great accurracy will not, of course, be obtained by using this simplified solution, or by using an alternative approximation adapted to solu-

Figure 3. Comparison between predicted and observed shelf-lives for packets of twenty cigarettes wrapped in "moisture-proof" regenerated cellulose film.
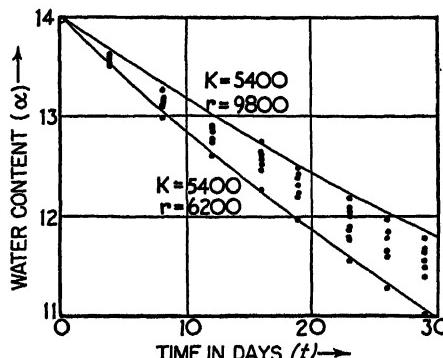
tion by slide rules.[26] But since storage conditions cannot be accurately foreseen, nor packages all made identical, it is better to use a simple approximation and a safety factor than to calculate the life exactly. For evaluating the majority of common protective packages, it is quite sufficient to use the simple equations based on the assumption of an exponential rate of decay.

REFERENCES

1. Halladay, J. F., *Paper Trade J.*, 115, T153 (1942); Oswin, C. R., *J. Soc. Chem. Ind.*, 62, 45 (1943).
2. Felt, C. E., et al., *Modern Packaging*, 20, 137 (1945).
3. Ministry of Supply, "The Use of Desiccants in Packaging," London, 1945; Sharman, C. F., Brit. Patent 582,746 (1945); Oswin, C. R., *J. Soc. Chem. Ind.*, 64, 67 (1945); Oswin, C. R., *Modern Packaging*, 20, 190 (1946).
4. Grover, D. W., *J. Soc. Chem. Ind.*, 66, 204 (1947).
5. Vogel, A. I., "Practical Organic Chemistry," p. 40, London, Longmans, 1948.
6. Gane, R., "U. K. Progress Reports on Dehydration, XI," London, H. M. Stationery Office, 1943; Makower, B., and Dehority, G. L., *Ind. Eng. Chem.*, 35, 193 (1943).
7. Graebner, W. H., Packaging Series No. 11, New York, Am. Management Assoc., 1944.
8. Oswin, C. R., *Chemistry and Industry*, 55, 183 (1949).
9. A.S.T.M., *Modern Packaging*, 22, 149 (1948); British Standard Specification No. 1133, p. 1619, London, 1948; Charch, W. H., and Scroggie, A. G., *Paper Trade J.*, 101, 201; Berry, W., *Plastics*, 33 (Jan. 1945).
10. Carson, F. T., Nat. Bur. Stand. Misc. Pub. M.127, Washington, 1937.
11. Anon., *World Paper Trade Rev.*, 397 (Feb. 8, 1946).
12. Hydon, W. L., *Ind. Eng. Chem.*, 21, 405 (1929).
13. Penn, W., *British Packer*, 10, 21 (1948).
14. Anon., *Modern Packaging*, 17, 91 (1943).

15. Parliman, J. H., *Modern Packaging*, **21**, 198 (1948).
16. Patton, C. W., *Am. Dyest. Reptr.*, **35**, 277 (1946).
17. Schaeffer, H. L., *Modern Packaging*, **20**, 150 (1947).
18. Edwards, J. D., and Strohm, D. B., *Modern Packaging*, **21**, 142 (1948).
19. Southwick, C. A., Jr., *Modern Packaging*, **18**, 142 (1945).
20. Oswin, C. R., *Chemistry and Industry*, **50**, 429 (1944); Pitman, A. L., *Food Ind.*, 49 (Jan., 1943).
21. Lavers, C. G., *Corrosion and Matl. Protection*, **4**, 6 (1947).
22. Thornthwaite, C. W., *Geographical Rev.*, **23**, 440 (1933).
23. Brooks, C. P., *Quart. J. Roy. Meteorol. Soc.*, **72**, 87 (1946).
24. Oswin, C. R., *J. Soc. Chem. Ind.*, **67**, 274 (1948).
25. Ellis, N. D., *Packaging Parade*, 40 (Sept., 1948).
26. British Cellophane Ltd., "The B.C.L. Method of Predicting Package Life," Bridgwater, 1946.

PACKAGING—AMERICAN PROBLEMS

Charles A. Southwick, Jr.

Technical Editor, "Modern Packaging," and Packaging Consultant

PACKAGING in the United States is considered a necessary element in the distribution, merchandising, and use of all products and goods. This concept has been greatly developed by the self-service type of retail store and from there it has spread even into industrial fields. This means that packaging does more than preservation alone, since the package is now the means of distribution and an aid to sales and merchandising. Actually, the concept of preservation is broadened to include the integrity of the count or weight, the guarantee of tamperproofness, as well as the preservation of quality and appearance.

Consumer Appeal. Furthermore, the package is important as a first sales aid by its attractiveness, by its presentation of a trade mark or design, or, if it can be seen, by the appeal of the product. The package becomes an aid to repeated sales by its maintenance of identity at the point of use, by its convenience in use, and by its delivery of a good product.

All of these elements are now used by companies who have broad markets and who wish to establish themselves with the final user, whether their products are metal parts, foods, drugs, cosmetics, chemicals, electrical goods, novelties, or any of the many types of products made and sold in our present economy.

New Materials and Methods

Today all industry is offering the benefits of its wartime accelerated research and development programs, and the result is new processes, machines, materials, and improved packages. Likewise, the package user whose wants have not been completely filled for the past many years is now interested in reducing costs, improving the appearance and function of his packages, and marketing new products. The net effect is that all forms of packaging are showing the results of these trends, particularly in the increased use of plastics materials.

Paper. Packaging still depends upon paper—alone, coated, laminated,

printed, and combined with other materials—to perform most of its tasks and functions. Processes for printing and coating as well as the manufacture of the base materials are being improved by the inclusion of plastic resins as sizings, adhesives, coatings, vehicles, modifying agents, and paraffin wax additives. The result is papers with more durability, better preservation factors, stronger heat seals, and better finishes for all purposes. By the use of adhesives which are wholly or in part made from synthetic resins, bags, fiber cans, fiber drums, shipping cases, multi-wall bags, and other strong paper containers have been greatly improved.

Coatings. The metal can industry has benefited by improved coatings which supplement tin, e.g., in beer cans, and allow a reduction in the use of this metal or eliminate its use entirely for many purposes. The glass industry is using more and more printing directly on the glass container, with resins or vitreous coatings which are attractive and as enduring as the glass itself; or it uses labels which are adhered to the glass by a special resinous coating which is capable of being activated by heat. The closures for glass containers are being improved by the inclusion of synthetic resins as coating materials for the liner or by the use of heavy sections of resilient plastics. Likewise, the cap itself can be made of an increasing number of different materials which are more decorative and more protective than those heretofore available. Flexible metal tubes are using synthetic resins and plastic coatings to increase their chemical resistance and durability of the printed exterior surfaces and to protect the interiors from corrosion. In many of the uses mentioned, the synthetic resins and chemicals lose their identity; and, while they are not striking examples of the uses of these materials, they should not be underestimated because of the large tonnage required to satisfy these requirements.

Plastics. However, nowhere is progress more noticeable than in the direct use of plastic or transparent materials to fabricate a package or to be its exterior surfaces. Plastic and transparent materials below 0.003-in thick are generally defined as "film," and those of greater thickness as "sheeting." The most popular film is, of course, cellophane, which continues to dominate the field and is finding expanding uses as carton over-wraps or as bag material, both printed and unprinted.

Next in importance is cellulose acetate, which is useful because of its dimensional stability towards humidity changes and so has found widespread use for printed bags and envelopes where moistureproofness is not required, or as laminated film over printed paper or metal foil for labels, wrappers, bags, and many other uses. Cellulose acetate as a bag or a wrapper is finding some applications for fresh green vegetables and fruits because of its ability to transmit readily most gases as well as water vapor.

The remaining films, such as those based upon rubber resins, other cellulose derivatives, vinylidene chloride, and polyethylene, are all finding increasing uses for particular applications where their physical or protective qualities are necessary. These uses include liquid type packages; carton, drum, and shipping bag liners for chemicals and abrasive materials; and many other highly specialized applications. In this group, polyethylene appears to be making the most rapid progress since it is low in cost and has outstanding properties and functions. In nearly every case, heat-sealing is the preferred method of fabricating these films.

In the heavier gauges, and particularly the transparent resins, there are increasing markets for rigid containers of various forms and shapes. Construction of these containers has been given impetus by the development of machines which turn them out automatically and at good speeds. Most of such containers are used for their decorative value.

Many synthetic resins are formed by injection or compression molding or by drawing processes into smaller packages of a multiplicity of colors, shapes and sizes for the packaging of jewelry, novelty items, cosmetics, drugs, and many other products which require an attractive, reusable, and durable package and where cost is not a limiting factor.

In the flexible packaging field, which includes bags, envelopes, folding cartons, and similar nonrigid package forms, there is a continuing trend towards improving the strength and efficiency of seams and seals; improving preservation factors, such as moistureproofness and greaseproofness; and towards making possible the packaging of liquids and the use of inert gases for preservation of fatty materials. All of these improvements are feasible because of new coatings, laminations of plastic films, and by improved fabricating and heat-sealing means which increase the package efficiency. An excellent example of a new and special requirement which was developed and became commercial is the use of a strong heat-sealable plastic film, as a bag for uncolored margarine. The film is a mixture of vinyl chloride and a Buna N type synthetic rubber which is heat-sealed into a bag. Held within this bag is a color capsule which, when broken by pinching from the outside, allows the product to be colored by kneading the bag and its contents. The plastic film's tensile strength and toughness tend to return the bag to a smooth shape after the coloring operation has been completed. This is a spectacular example but one which is indicative of the capabilities of plastic films when fabricated by heat-sealing means, and which allows full advantage to be taken of the film properties.

Improved Packaging Machinery. Supplementing the new materials are many improvements in machinery and the means of producing packages. The present trends in packaging machinery are towards higher

speeds, improved efficiency, the inclusion of electronic devices, and the ability to handle the newer resins and films. However, as a result of a continuing trend toward single use or smaller size of packages with improved preservation qualities, many machines are being made to form, fill, and seal the package. There appears to be a continuation of the wartime interest in packaging for the preservation of metal parts of all kinds, with added interest in combining preservation with a simplified means of inventory, display, and merchandising for precision and nonprecision metal parts of all types. For this, synthetic films such as polyethylene appear to be ideally suited and have been finding increasing acceptance.

Food Packaging. Meanwhile, a great deal of development and research continues to find improved, simplified, and automatic means for handling frozen food packages, and to perfect methods of packaging fresh fruits and vegetables so that they may be brought to the consumer trademarked at the peak of ripeness and on a 100 per cent usable basis. However, this problem involves more than packaging since production, distribution, and transportation will have to be integrated to achieve the final result. The problem is complicated by the varied metabolic requirements of fruit and vegetables, and at present it is not possible to package and preserve many of these foodstuffs at the farm. Today produce packing is extensively practiced for convenience and display rather than for preservation, and it is usually done not too far from the point of sale.

Conclusion

Along with improvements in machines, packages, and materials, there has been a growing recognition of the need for a scientific and engineering basis of packaging. The expanding use of synthetic resins and many of the other products of the chemical industry, the complexity and high level of functions required by many of today's products, have made it necessary for scientific and engineering methods to be used in the development, evaluation, and production of packages. In view of the present situation and trends in this country, packaging should be redefined so that it may now be stated that: "Packaging is a low cost, convenient, and protective means of distributing and merchandising a product."

REFERENCES

1. "Modern Packaging Encyclopedia," New York, 1950, articles on "Developing the Package," pp. 92-105; "Plastic Packaging Films," pp. 177-182; "Marking Glass Containers," pp. 442-444; "Choice and Use of Adhesives," pp. 215-221; "Plastic Packages," pp. 505-508.
2. Southwick, C. A., Jr., "Packaging Uses of Polyethylene, Packaging Series No. 29, New York, Am. Management Assoc., 1949.
3. "Thermosetting Enamels for Glass," *Modern Packaging*, 190 (April, 1949).
4. "Vinyl-Nitrile Blends," *Modern Packaging*, 149 (December, 1948).

FREEZE-DRYING *

Earl W. Flosdorf

Director Research and Development, F. J. Stokes Machine Co., Philadelphia, Pa.

IN FREEZE-DRYING, the process of drying by sublimation, the product is first frozen then sublimed under high vacuum—that is, the ice evaporates without melting. The solids contained in the original product retain the shape of the original frozen product and there is little overall reduction in volume. The water vapor in the high vacuum is continually removed by any one of three methods: a cold condenser may freeze out the water vapor in the form of ice; a chemical desiccant may remove it by chemical reaction; or a vacuum pump may pump all of it out into the atmosphere. Freeze-drying can take place without a vacuum, but the process would take far too long to be useful. By means of a vacuum pump, the air molecules are removed so that the sublimed water vapor molecules can travel rapidly to the cold condenser, the desiccant, or through the pump. Where a vacuum pump only is used (without cold condenser or desiccant), it must be sufficiently large to remove all the water vapor as well as air.

The method chosen for the removal of water vapor depends upon many factors of specific application, e.g., whether laboratory scale of operation or full production, the type of product being dried, and the like. For large installations, a cold condenser is usually best. In this case, the condenser is cooled by ammonia or "Freon" refrigeration. In some smaller installations, "Dry Ice" may be used to cool the condenser. Cryochem † equipment utilizes a desiccant and is useful in laboratory operation; moreover it can be operated in far-away and remote locations such as China, India and Africa where "Dry Ice" is unavailable.

The temperature at which the products must be dried is, of course, below freezing; how much below freezing depends entirely upon the nature of the specific product. In any case, heat is added to compensate for the loss of latent heat of evaporation and must be supplied rapidly

---

* E. W. Flosdorf is the author of "Freeze-Drying: Drying by Sublimation," New York, Reinhold Publishing Corp., 1949.

† From the Greek *kryo* = "cold," "frost," and *chem* from "alchemy."

665

enough to speed the process without causing the material to melt. Average temperatures for drying are in the range of −15 to −35°C. This corresponds to a range of vapor pressures from about 180 to 1300 microns on a McLeod gauge. The cold condenser, chemical desiccant, or vacuum pump must provide a vapor pressure lower than that corresponding to the ice being sublimed in order to create a driving force.

Drying by this process avoids the danger of destroying the desired therapeutic properties of serum, blood plasma, penicillin and other biological products. In addition, many chemicals and pharmaceuticals are dried this way in order to improve their rate of solubility. Considerable research has been carried out on sublimation drying of a variety of commercial materials—including fruit juices, milk, fish and meat—and the process and equipment have been streamlined to the extent where the resulting products may soon appear on the market.

Advantages of Freeze-Drying

The specific advantages of freeze-drying are numerous, and explain the widespread application of this method in industry.

(1) The temperature is below that at which many labile substances undergo chemical change. This applies to labile components in blood, to viruses and most forms of microorganisms and to other biologicals and pharmaceuticals.

(2) Because of the low temperature, the loss of volatile constituents is minimal, which is particularly important with many foods like orange juice and pineapple juice.

(3) Since the product is frozen, there is no bubbling or foaming, which in some cases causes changes due to surface action, e.g., the surface denaturation of proteins which occurs in drying their solutions under vacuum, even at low liquid temperatures.

(4) As the frozen solvent sublimes the solute remains evenly dispersed and distributed without undergoing concentration, and the remaining dry residue emerges as a highly porous solid framework. It occupies essentially the same total space as the original solution did and the final residue is not the fine powder with which the chemist is most familiar. It consists of a friable, interlocking, and sponge-like structure. As a result, solubility is extremely rapid and complete. For example, gelatin dried from a solution which had to be prepared in the first place by boiling becomes instantly soluble in cold water.

(5) Since the molecules of solute are virtually "locked" in position in this way, the tendency for coagulation of even lyophobic sols is minimal. Even though the lipoidal constituents of dry blood plasma do not reconstitute perfectly after drying and produce a slight degree of turbidity,

there is far from complete coalescence. The particles are small enough to be safe for intravenous injection and do not cause capillary embolism.

(6) During drying, the surface of the evaporating frozen ice layer gradually recedes to leave more and more of the highly porous residue of solute exposed. As a result, "case-hardening" never occurs. A far lower content of moisture may be obtained in the final product without the use of excessively high final temperature. Because of this lower moisture content, a greater degree of stability results than with any other method of drying.

(7) Bacteriological growth and enzymatic changes cannot take place under conditions of freeze-drying. This is important for foods as well as medical products used in parenteral injection. The final fully dried product likewise resists bacterial growth and enzymatic action.

(8) Because of the high vacuum used, in contrast with the degree of pressure used in ordinary low-temperature liquid evaporation, the amount of oxygen present is so extremely small that even the most readily oxidizable constituents are protected.

For these reasons listed above, freeze-drying has become a major drying process in the biological and pharmaceutical industry. The detailed history of the development has been recorded, and a complete bibliography given.[1] Long ago, substances were dried in an ice-box, and early in this century this procedure was combined with drying in a vacuum chamber. However, within the last ten or fifteen years it has been recognized that the process could be applied industrially and improved products would result by the establishment of proper vacuum conditions for the removal of water vapor and by the rapid application of heat to the frozen product, rather than by keeping it in an ice-box. The final dried product is raised to a temperature well above 0°C, and often as high as 60 or 70°C, in order to reduce the final content of moisture to a minimum.

Two Stages of Drying

There are two stages in drying by sublimation. In the first, ice is evaporated from a frozen mass. In the second, moisture is removed from the final dry solid to lower the residual content to a minimal level. During the first stage, depending upon the particular product, about 98 to 99 per cent of all water is removed. In the second, the residual moisture content is reduced to 0.5 per cent of the final product or less, which represents final removal of 99.95 per cent of the original content of water (assuming 10 per cent solids originally). In the first stage, temperatures are well below 0°C. As the process gradually passes into the second stage, the temperature rises and finally reaches that of the room or higher, according to the heating means used.

Containers for Medical Products

Medical products are generally dried in the final containers used for market distribution, although this increases the cost of drying. Wherever the products are used for parenteral injection, this is a safer procedure, because representative bacteriological sampling can be carried out more accurately with liquids than with solids. Accordingly, when the final containers are filled with liquid and sampled for test before freezing and drying, there is greater certainty that products and containers are sterile than when sampling is done after the dried product has been weighed and distributed to the containers, although the latter can be done.

Bulk Drying

For many purposes, it is desirable not to dry in the final ampoule. This is the case in carrying out desiccation of products as part of a general processing in which the final product is not stored or distributed in dry form, as, for example, serum albumin. Similarly, labile substances may be concentrated in this fashion for other purposes. Either special stainless steel sterility pans with lids designed to keep out microorganisms, or large glass or metal bottles, may be used. Bottles may be placed either on a manifold or within heated vacuum chambers. Foods, of necessity, must be dried in bulk.

Rapid Drying Cycles

It is self-evident that water vapor must be carried off quickly in order to effect rapid drying. By means of modern apparatus this is readily accomplished. An adequate condenser with proper surface temperature to form ice, or a steam ejector designed for correct pressures (in either case, having sufficient capacity in relation to the volume of products to be dried), can be specified for as favorable a ratio as is desired. On the basis of removal of water vapor alone, there is no physical limitation with regard to speed of drying. The controlling factor is one of rate of supply of heat for rapid evaporation in order to maintain the vapor pressure of the frozen product at the optimal point.

Drying and Storage Changes

Changes in desiccated products may be considered on the basis of chemical and physical alteration, immunological change, loss of viability of living organisms, and loss of other characteristics of biological activity. These are considered in detail elsewhere.[1]

Of the various chemical substances that have been studied following desiccation from the frozen state, perhaps proteins have received the most attention. Generally, no chemical alteration detectable by ordinary methods occurs during drying of these substances. Electrophoretic patterns of plasma dried by sublimation have presented a normal appearance with normal values. If the proper degree of dryness has been attained, changes do not occur during subsequent storage for reasonable periods. Fats show no chemical change during drying, but undergo oxidation on storage and tend to become rancid if oxygen is not entirely excluded, even when the product is completely and properly desiccated and kept in a moisture-free condition. Hydrolysis also occurs slowly. Other chemical substances have not been studied extensively, but in general this method of desiccation is as gentle as any and is far less likely to produce changes in labile substances. Because the final moisture content can be reduced to such a low level without harm to the product, greater stability of dry product is obtained than by any other method. Reduction to such a low content of moisture by other methods usually results in loss of activity or solubility or other characteristics, and is therefore not practicable.

Physical alteration tends to occur rather readily in the case of lyophobic sols as a result of desiccation. This accounts for the precipitation of lipoidal substances referred to previously in connection with human serum and plasma. It occurs to a lesser degree with rabbit-serum and is not noticed with horse-serum. Likewise many bacterial suspensions, unless adequately protected with a colloidal protective agent, such as milk, do not resuspend properly when reconstituted. Other substances, such as lyophilic protein, have remarkable properties of rapid and complete solubility, even after storage for years. The solubility characteristics of many products are increased remarkably as a result of drying by sublimation. Most notable of these are gelatin and vitamin B preparations.

Immunologically, little change in either serum or antigens is observed. Crystallized egg-albumin has been kept, after desiccation of aqueous solutions from the frozen state, for four and five years without noticeable change in antigenic specificity. Immune sera undergo no detectable loss in titer following processing or storage for years at room temperature. Perhaps as sensitive an index as any is the complement activity of guinea-pig serum. Complete and proper drying of this product not only results in no loss of activity but also maintains the complement in stable form for a period of years. By the addition of a stabilizing solution, consisting of sodium acetate and boric acid, to guinea-pig serum before drying, a product is obtained which keeps well and which is stable without apparent loss of potency, for a period of several days following restoration.

In small hospitals and laboratories, this factor adds to the convenience of use of desiccated complement.

There is a loss of CO_2 during processing of serum, but the resulting increase in pH has no effect on its complement activity. Restoration with acid or buffer is unnecessary. The alkalinity of reconstituted human serum or plasma does not cause reactions; it is unrelated to turbidity. As indicated above, the lyophobic lipoidal constituents are the cause of turbidity and do not render the product unfit for injection.
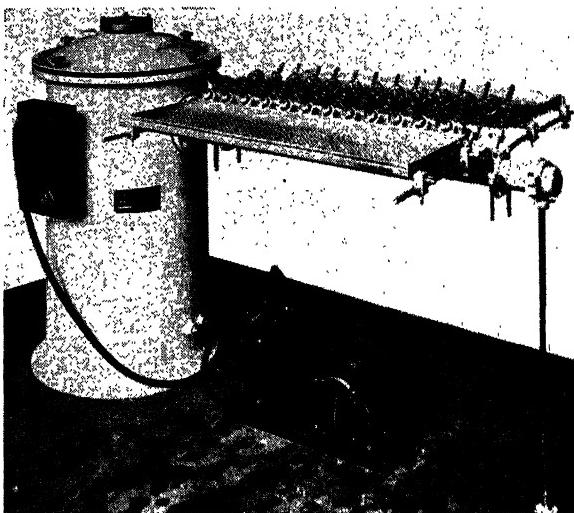
Figure 1. Cryochem-type equipment for desiccating from the frozen state. A regenerable desiccant comprising mostly calcium sulfate (Drierite) with a small amount of silica gel (for a small amount of vapors other than water) contained in baskets within the white tank. The manifold is shown with outlets to which individual containers for drying are attached. At the end of drying, the glass tubes attaching the containers to these outlets are sealed by fusion with a flame.

Turbidity, however, does cause difficulty in carrying out precipitation reactions, as with rabbit-precipitin or in diagnostic tests for syphilis. This has been a drawback to the use of dry standardized syphilitic human serums for calibration or standardization of diagnostic test-procedures. These sera could facilitate standardization of test results in different laboratories. The difficulty with lipoids may be overcome by double processing by means of intermediate Seitz filtration. At the time of filtration of the reconstituted serum, following the primary drying, the titer of the serum may be adjusted by dilution to a desired point. It is then dried for the second time. The final dry product accordingly has a predetermined definite titer.

In the case of many viruses, there is a loss in activity as a result of processing, but the final desiccated product is well stabilized and no further loss occurs during storage. Viruses are in general more difficult to dry than other substances. The temperature must be maintained quite low during desiccation, preferably well below −20°C. The percentage of viable cells remaining in cultures of bacteria after desiccation from the frozen state is as low as 5 per cent, but the surviving cells keep well. The initial survival may be materially increased by using a protective protein, such as milk. The biological activity of many other substances, such as enzymes, penicillin, certain hormones and vitamins, and other labile substances is maintained without loss both during drying and subsequent storage.

Equipment

For laboratory work, the capacity required is usually small and either "Dry-Ice" low-temperature condensation or chemical methods for removal of water may be used. Bottles of material may be placed on a manifold or in a drying chamber. Figure 1 illustrates one type of manifold with
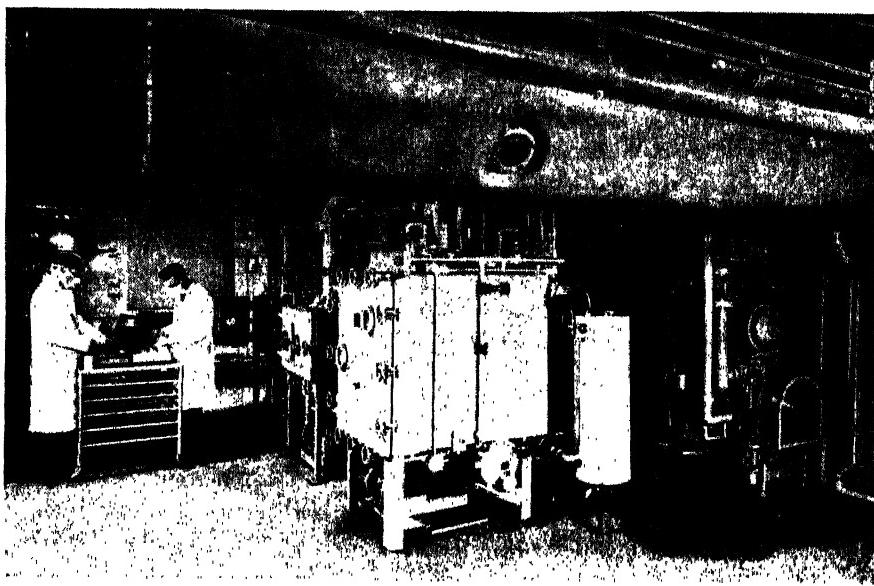
Figure 2. Typical medium-size plant for blood plasma and similar products. (See text for description.)

supporting racks for containers; the manifold is attached to a cryochem chamber containing regenerable chemical desiccant. When drying chambers are used, they must be equipped with some means of heating because of the vacuum insulation of the containers from the atmosphere.

For larger-scale operation producing blood plasma and similar products, drying chambers are more satisfactory. They can be used in conjunction with any one of several methods for the removal of water vapor. The use of "Dry-Ice," however, will generally be found to be unnecessarily expensive. A well-designed type of mechanically refrigerated low-temperature condenser is recommended, either "Freon" or ammonia refrigeration being used. A plant of this type is illustrated in Figure 2. Two different size drying chambers are illustrated, with the hot-water circulating system for heating. The large vapor condenser can be seen in the foreground suspended from the ceiling. Vacuum pumps are shown at the extreme right. The refrigeration compressors are located in another room.

Full consideration of equipment for commercial-scale production and engineering details are discussed elsewhere.[1] In brief, drying chambers of suitable size are used, with steam ejectors or mechanically refrigerated low-temperature condensers. Steam ejectors must be either four- or five-stage. The choice between steam ejectors and mechanically refrigerated condensers is made on the basis of the relative cost and availability of electricity for refrigeration, and of high-pressure steam and low-temperature cooling water for the ejectors. Large quantities of steam and water are required for the latter. Processing with either type of equipment is convenient and gives the same type of product when proper conditions are set up.

Conclusion

Fifteen years ago, the freeze-drying process as such was unknown, and prior to that time it was a laboratory curiosity. Today, there are hundreds of laboratory installations, including "homemade" and specially designed equipment. Almost every major pharmaceutical manufacturer utilizes the process for some commercial operation. This serves as an index of the rapid progress which has been made in the development of freeze-drying.

REFERENCE

1. Flosdorf, E. W., "Freeze-Drying," New York, Reinhold Publishing Corp., 1949.

SOME PHYSICOCHEMICAL ASPECTS OF MARGARINE MANUFACTURE

H. W. Vahlteich

V. Pres. of Research and Quality Control, Best Foods, Inc., New York, N. Y.

IN A SMOOTHLY OPERATING large-scale margarine manufacturing operation there lie hidden a host of variables, nicely and smoothly adjusted to each other and unnoticed by the casual observer. This is true of many large-scale operations and tends to belie the tremendous amount of study and research conducted by the operators and the technical staffs over a period of years. No well-working margarine manufacturing operation comes suddenly into being. Bacteriologists, chemists of various trainings, engineers and experienced operators are all links in the chain. To say that any one specialist is more than partly responsible for such an operation is fallacious, but it can be said with confidence that the physical chemistry involved is fundamental to most phases of the operation.

Early Developments

To appreciate the problems of manufacture it is desirable, and even necessary, to approach the subject historically, as well as from its chemical and engineering aspects. Margarine manufacture has now reached the ripe old age of 80, reckoned from Megé-Mouriez' early attempts to make the product in the late 1860's. These and the history of margarine manufacture before 1880 have been cited so frequently[1-5] that simple reference to them will satisfy the reader interested in modern margarine manufacture. Generally speaking, for some time prior to Storch's[6] early work on the culturing of milk around 1890, nothing of any great importance took place to influence modern margarine manufacture. This statement is not to be taken too literally because by 1886, at least in the United States, margarine manufacture had attained sufficient importance to be influenced by political events. In that year, in the United States, there was enacted by Congress the Oleomargarine Act of 1886, the first of the discriminatory anti-margarine laws, which at least indicates that

the volume of margarine sold in the United States, although small by present standards, was sufficient to attract political attention.

Storch's work on the fermentation (culturing) of milk for butter manufacture did, however, have a tremendous influence on margarine technology because it introduced bacteriological control of flavor development. This phase of margarine manufacture has merited and received considerable attention, e.g., the work of Hammer and his associates at Iowa,[7] to be followed in the late 1920's by the work on the flavor ketones by van Niel et al.,[8] Schmalfuss,[9] and many others.

Chronologically, there appeared about 1900 another development,[10] namely the introduction of the use of egg yolk as an emulsifying agent, which is of importance for a number of reasons. Egg yolk was presumably used simply to keep margarine from spattering in the frying pan. It was, therefore, the forerunner of vegetable-origin lecithin, of which more will be said later. Egg yolk never proved very popular for the above purpose because of its high cost and because it rendered the margarine containing it more susceptible to bacterial inroads than the product not containing it. Nor were the more refined egg-yolk products, such as that described by Steudel and others,[11] sufficiently superior in these respects to replace egg yolk permanently.

Hydrogenation and Deodorization

Going back now to the turn of the century, there were two technological developments affecting the principal constituent of margarine, namely margarine oil.* These were the solidification of limpid oils to any desired melting point by hydrogenation and the development of "deodorization." The former resulted from the scientific discovery by Sabatier and Senderens [12] that unsaturated organic compounds could be hydrogenated in the presence of a catalyst and the application of this discovery by Normann [13] to glyceridic oils. Deodorization of glyceridic oils evolved somewhat gradually, and rather few references are to be found in the scientific and patent literature on the details of its evolution. These developments are of such great significance in margarine manufacture— its physicochemical and other aspects—that it is well worth the time to consider them in some detail.

Weber and Alsberg [14] give an interesting account of the impact of hydrogenation of glyceridic oils on the entire industry. They point out that prior to the advent of commercial hydrogenation, the amount of solid fat (edible) products which could be manufactured was dependent upon normally solid animal-fat products to a large degree. These were

* Referred to in the industry as margarine oil rather than margarine fat, probably because it is maintained molten for ease of handling until the final stages of margarine manufacture.

primarily tallow and the solid fractions of grained and pressed lard. The liquid fraction of lard or lard oil had a variety of uses (such as for illuminating oil and greases) which were being replaced by petroleum products and electricity (as well as by illuminating gas), so that the hard fraction of pressed lard became increasingly expensive as the uses for lard oil passed out of existence. It was in such a tense economic atmosphere that the prospect of manufacture of hard fat, in almost infinite variety of hardness, from liquid vegetable oils was suggested. It was well too, that this was a major problem and its solution attended by heavy economic pressure, because the successful commercial hydrogenation of glyceridic oils required not only the engineering of new processing equipment, but also a vast amount of scientific and technological research to perfect a satisfactory catalyzer and to attain adequate sources of hydrogen gas. Indeed, although Normann's patent was issued in 1902, commercial hydrogenation of glyceridic oils was not well on its way for another ten or a dozen years, or just prior to World War I.

Even so, as late as the early 1920's, so-called "selective" hydrogenation was regarded as being of a single type: the selection of the multi-double bonds to be hydrogenated to oleic acid (the single double bond) before the oleic acid double bond was hydrogenated to stearic acid. The concept of selectivity in hydrogenation which would yield a maximum of iso-olein, in preference to olein, and its application to margarine manufacture did not materialize for perhaps another ten years, as will be reviewed later.

Weber and Alsberg [14] also give an interesting historical account of the development of deodorization. From their account, it would appear that the origin of the process goes back, possibly 100 years, to steam blowing of a variety of fats and fatty acids, gradually increasing temperatures being used. Since steam is an inert gas in this operation, it served well until applied to edible fats at the temperatures used. This use of higher temperatures, plus the dissolved air in glyceridic oils, probably explains why a none too satisfactory result was attained in rendering edible oils palatable. With the advent of high-vacuum (below 10 mm), steam deodorization, a degree of perfection in palatability (and stability) was attained which made possible the flavor acceptance of edible fat products, quite beyond anything accomplished theretofore.

Broadly speaking then, the advent of hydrogenation of glyceridic oils and of their high-temperature, vacuum steam deodorization were well nigh simultaneous in the first two or three decades of this century. They were probably the two greatest single developments in the progress of edible fat technology, although not the only ones. Other changes in margarine processing, included the advent and later near-disappearance of coconut oil; the newer techniques of hydrogenation; the development of

edible emulsifying agents somewhat belatedly, following the Harkins-Langmuir [15] concepts of interfacial activity; as well as the purely engineering developments in packaging equipment and internal chilling units for solidifying the margarine emulsion. These, which have been mentioned in their approximate chronological occurrence, and several others have a direct bearing on the physical chemistry of margarine manufacture.

Until the early years of the twentieth century, oleo oil and neutral lard were widely used in margarine manufacture. Just prior to World War I, coconut oil came into use for margarine fat because of its increasing availability, the fact that it could be hydrogenated and deodorized very successfully, and because of its physical characteristics. Since about the mid-1930's, it has largely disappeared as a margarine fat. This was due to its political disadvantages (in the United States) and the fact that its hardness below about 55°F became an increasing drawback with the institution of domestic mechanical refrigeration. Its political disadvantage arose from the fact that American farmers were competing with foreign fats producers at a time when such competition was supposedly injurious to the (American) farmer. Its plasticity range was a challenge to the margarine technologist for margarine is a product intended to be used as a spread for bread, and when made from coconut oil it was too hard to spread under 55°F (below which temperature mechanical refrigerators normally operate). Also, it became practically liquid at 75 to 80°F, and so never maintained its shape in warm weather.

The experience of margarine technologists with coconut oil, however, led to a single discovery from which stemmed, to a major degree, the development of modern margarine. Coconut oil came into use, with the advent of hydrogenation, not alone for margarine, but as a confectioner's fat. Generally speaking, edible fats may be divided into two groups, based on their physical properties, the laurin-myristin group and the olein-palmitin-stearin group. The physical properties of these two groups of fats are strikingly different, and hence they are not always easily interchangeable for specific uses. The laurin-myristin type includes such fats as coconut, palm kernel, babassu, tucum kernel, and muru muru oils — all tropical fats. The olein-palmitin-stearin type includes the animal fats and the domestic (U.S.) edible vegetable oils, primarily cottonseed, soya bean, peanut, and corn oils.

While these two types of fats are interchangeable from the standpoint of the calories they furnish, they do possess characteristic differences which definitely interfere with their interchangeability for specialized uses. Thus, the laurin-myristin fats are characterized by their "quick melting," their "cooling-effect-on-the-tongue," and their "dryness." They are also regarded as "less oily" than the fats of the other group. Davis [16]

showed that coconut oil was a relatively poor shortening compared to lard. Indeed, all these curiously expressed properties point to the laurin-myristin fats as more desirable for flavoring uses, and hence it is not surprising that their principal edible uses are in the light bakery and confectionery trades. The olein-palmitin-stearin fats, on the other hand, are the cooking and shortening fats, generally characterized by being plastic over a much wider temperature range, by higher melting points for the solid fats, by a more noticeable oiliness, and by superior shortening qualities.

To go much beyond these very general characterizations leads to difficulties, however, because there are individual oils and fats in each group which have important properties belonging in the other group and, in addition, many fats and oils have characteristic flavors which cast their physical properties in a subordinate role. For example, cocoa butter is definitely a confectioner's fat but is not in the laurin-myristin group, and peanut oil is chemically in the olein-palmitin-stearin group but has characteristics which favor its use in the light bakery and confectionery trades. Again, olive oil has a flavor which creates for it a specific demand, soya bean oil has a flavor which is also characteristic and difficult to remove, and animal fats usually have a characteristic flavor which is as much liked by some as it is disliked by others.

Cocoa butter has long been the confectioners' fat par excellence. Its properties dominate the salability of products of which it is a part. Its strong points are its fine characteristic flavor and the narrow spread between its setting and melting points, whereas its weak feature is its relatively low melting point. The latter is so crucial that cocoa butter cannot be used in candies and coatings in warm summer weather, a detail which led to a long series of trial and error investigations, first to improve it, and then to replace it by other fats. The separation of cocoa butter into soft and hard fractions by hydraulic pressing met with only minor success. Hydrogenation cannot be applied to cocoa butter because it does not lend itself readily to subsequent deodorization. The hydraulic pressing of the laurin-myristin type of fats did, however, produce a "cake" or hard fraction which made an acceptable substitute for cocoa butter and especially so after hardening by hydrogenation. Coconut oil can be very successfully deodorized and coconut-oil cake had many of the desirable flavor and physical properties of cocoa butter, such as "quick melting," "cooling-effect-on-the-tongue," and "dryness." It was noticed many years ago that cocoa butter had a congealing or setting point much closer to its melting point than most other fats in common use at the time. In this respect coconut-oil cake resembles it while other hard fats, such as hydrogenated cottonseed oil, do not. Although this was generally known in the early 1920's, it seems that not much, if any, practical application had

been made of this fact. However, it was then generally known that while the hydrogenation of coconut-oil cake raised its melting point noticeably, it was not thereby materially improved in "dryness"—that is, lack of oiliness—at summer temperatures.

The principal disadvantage of pressing coconut oil lay in the yield of cake, which was only about 25 per cent, and the difficulty in getting a good price for the 75 per cent of soft fraction. Then someone discovered that palm-kernel oil could be pressed like coconut oil, but with a higher yield (35 to 40 per cent) of a much better cake. Although the yield was improved considerably, a far more important development arose from the observation that palm kernel-oil cake was much "drier" than coconut-oil cake and much superior to cocoa butter as a coatings fat. Then came the observation that this improvement in "dryness" paralleled the setting point and not the melting point of the fat. The history of this development has been relatively simple since that observation was made—merely a matter of finding a laurin-myristin type of fat which had a suitable setting point, and which did not require any pressing operation. This was achieved in two steps by means of two palm-type oils of the botanical family *Astrocaryum*, namely, tucum kernel oil, which yields 65 to 70 per cent of suitable cake upon pressing, and more recently, muru muru oil, which requires no pressing but has a suitable setting point and "dryness" as commercially produced. Both of these oils are somewhat improved by hydrogenation and are readily deodorized to a perfectly neutral flavor. Both have come to this country from Brazil in recent years.

The simple discovery referred to above was that the setting point (also referred to as the congealing point) of a fat, and not its melting point, is the key to its ability to withstand warm weather temperatures; and that it is essential that its melting point (Wiley) is sufficiently below body temperature (98.6°F) to assure rapid melting (in the mouth) of the margarine from which it is made. Very briefly stated, the problem is to attain a melting point of 93 to 95°F with a setting point of 77 to 79°F. This setting-point range will vary somewhat, but not significantly, depending upon details of hydrogenation procedure. It goes almost without saying that these melting-point and setting-point ranges are rather narrow and call for considerable skill in hydrogenation as well as for careful laboratory control.

There are two more-or-less classic ideas about the hydrogenation of glyceridic oils. The oldest and simplest involved the idea that olein, with one double bond, was hydrogenated to stearin by the addition of hydrogen at the double bond. Then came the somewhat more complicated conception of "selective" hydrogenation which involved the idea of so hydrogenating a mixture of olein and linolein as to hydrogenate first most or all of the linolein to olein before hydrogenating the olein to

stearin. This second concept was well-established in the shortening in-
dustry in the 1920's. About that time there also began to appear evidence
of "new acids of hydrogenation"—i.e., fatty acids of glycerides produced
during and characteristic of hydrogenation; also sometimes referred to as
isooleic acid. It was discovered that a cottonseed oil, for example, when
hydrogenated in such a manner as to produce a suitable margarine fat—
that is, a margarine fat having a melting point well below body tem-
perature with a setting point of about 77°F—must be treated so as to
produce a maximum of isoolein and that this was not only dependent
upon the conditions of hydrogenation and the catalyst, but also upon
the amount of linolein in the starting oil. The higher the linolein con-
tent of the oil to be hydrogenated, the more easily was an adequately
high setting point achieved without exceeding a melting point of over
about 95°F. This leads to another concept of the term "selective" in
referring to such hydrogenation—namely, that the hydrogenation can be
controlled not only selectively to hydrogenate the linolein first to olein
and then to stearin, but also to hydrogenate the linolein first to isoolein,
and then possibly to olein, before it is further hydrogenated to stearin.
Since isoolein is solid at temperatures at which olein is liquid, such
selective hydrogenation produces more glycerides solid at room tempera-
ture than if an equivalent reduction in iodine number has been achieved
by converting linolein to olein as is done in hydrogenating shortenings.
This, briefly, is an explanation of how and why it is possible to produce
a margarine fat of satisfactory physical constants from any one or more
of the four edible domestic vegetable oils available in large quantities in
the United States—namely, cottonseed, soya bean, peanut, and corn oils.

In addition to satisfactory performance in warm weather, a margarine
made from domestic vegetable oils is easily spreadable at 50°F, and so
offers no difficulties on this score when removed from the refrigerator.
In fact, it is more spreadable at temperatures below 50°F than are butters.
In addition, it is entirely the product of the American farm, which aids
materially in solving some of the political problems involved.

Essentially it is necessary to hydrogenate "domestic" vegetable oils
cottonseed oil, soya bean oil, peanut oil, and corn oil in a manner dif-
ferent from that used for hydrogenating these oils for shortening; that is,
it is the opposite type of hydrogenation. This is the case because in oil
hydrogenation to produce shortening it is desirable to reduce the "iso-
oleic acids" (their glycerides) to a minimum, while in margarine-oil pro-
duction it is desirable to achieve a maximum of these glycerides. For this
purpose too, opposites are chosen in the three general fundamentals
applied to these hydrogenations, namely: (1) hydrogenation processing
conditions, (2) production of a catalyst of suitable characteristics, and
(3) the selection of suitable oils for starting stocks.

Emulsification

Chronologically, another involved physicochemical development, and one of very practical commercial significance, somewhat preceded the advent of the newer hydrogenation techniques which made it possible to produce an improved margarine from limpid vegetable oils. While this development started earlier, it grew to practical fruition more slowly than hydrogenation. It was the development of a concept of emulsification, of the play of interfacial forces, and the practical steps necessary to control and eliminate certain undesirable features attendant upon margarine manufacture and distribution. Anyone who can recall margarine in the United States 20 to 30 years ago [17] knows only too well the expression "weeping" as applied to margarine. This term was descriptive of the inability of the finished product, in the package, to retain its milk moisture, especially when made from coconut oil or other similar laurin-myristin type oils. The all too frequent result was a moisture-soaked and dripping package which led to consumer complaints as well as to short weight troubles. It also touched deeply upon manufacturing procedures because, although coconut-oil margarine could reasonably be expected to retain only 12 to 14 per cent milk moisture, something like twice this amount of milk moisture had to be worked out of the chilled margarine emulsion before packaging it as margarine. Also, the finished product usually retained less than half the table salt used in its manufacture.

It became important, then, to change the manufacture of margarine from processing which involved the removal of one-third to one-half of the starting materials and water entrapped in the process, to a starting formula which did not eliminate any ingredients and a processing procedure which added none to be removed. In turn, this meant the elimination of the old "wet-chilling" procedure and the development of emulsifying agents which would prevent weeping. The scientific background for our modern concepts on emulsification was set by Harkins and by Langmuir,[18-22] and is described by Harkins in his chapter on surface energy in colloid systems in Bogue's book.[16] An apparently empirical discovery of considerable practical value, involving this concept, was published by Luksch in 1922,[23] who reported that mono- and diglycerides were very effective moisture-retaining agents when used in margarine to the extent of only 0.1 to 0.3 per cent. Luksch's article was brief and did not relate his reported discovery to the Harkins-Langmuir surface-energy concepts except in so far as he pointed out that the effectiveness of these agents was apparently proportional to their monoglyceride content as distinct from their diglyceride content. This has been adequately confirmed and hence it is probably safe to hypothesize that the emulsification value anti-weeping effect of monoglyceride prepara-

tions is proportional to the number of OH groups in the molecule. This, in turn, would mean that the moisture-retaining capacity of mono- and diglycerides is proportional to the amount of water-soluble (OH) groups in the molecule.

Meanwhile, in the middle and late 1920's and early 1930's, more experimental evidence was being reported in the literature to build up the practical value of the Harkins-Langmuir concept as it applied to margarine manufacture.[24] From about 1925 to 1930, soya bean oil began to be produced in sufficiently large commercial quantities and, due to its relatively high lecithin content, resulted in the commercial availability of soya-bean lecithin.* In the course of a few years, the relatively low price of this lecithin made it available for use as an emulsifying agent in margarine. This readily oil-dispersible and easily hydrated compound has a special usefulness in margarine quite distinct from that of monoglycerides. A quantity of $\frac{1}{8}$ to $\frac{1}{4}$ per cent of commercial soya-bean lecithin (60 to 65 per cent total phosphatides) in margarine gives the latter very desirable frying properties by eliminating spattering in the frying pan.

It was soon discovered by margarine technologists that these two newly available emulsifying agents for use in margarine had separate and distinct effects in the finished products. Lecithin had little or no capacity to prevent "weeping," while monoglycerides did not prevent serious adhesion of milk solids to the frying pan and only moderately reduced spattering.

Rosenbusch and Reverey [25] published a procedure for estimating quantitatively the anti-spattering effect of an emulsifying agent intended to reduce spattering in margarine, and the estimation of the effectiveness of an anti-weeping agent is easily achieved by simple observation of the absence of weeping and the moisture content of the margarine.

Research in this field also resulted in the development (Harris [24]) of the sodium sulfoacetate derivative of monostearin, which has been successfully used to achieve the combined effects of milk moisture retention and the reduction of spattering in the frying pan. Harris, in a series of patents, has also disclosed a number of other compounds serving either one or both of these purposes, requiring only small amounts in margarine. Many more of such compounds are and will in the future be available, since the foundation for their development has been clearly outlined. It remains then, for the organic chemist to synthesize harmless compounds with a variety of combinations of hydrophylic and lyophylic properties. A series of compounds, the monocitrates, recently have been suggested for some of these purposes.[26] Curiously enough, there has never

---

* See paper by J. Stanley in Vol. VI of this series.—*Ed.*

been developed a method of scientific prediction on the basis of which a compound with the desired anti-weeping and/or anti-spattering properties could be worked out on paper, with the assurance that, once synthesized, it would be practically operative in margarine.

By means of the newer emulsifying agents developed in the past 15 to 20 years then, it is possible not only to produce a margarine which behaves well in the frying pan, but also it is possible to start processing with all the ingredients in the churn that will be in the finished product. This eliminates not only the waste of milk and salt, but also the cumbersome bulky equipment and space formerly required. The external and internal dry-cooling of the margarine emulsion have practically eliminated the chill vat and similar wet-chilling procedures.

Cultured Milk

While margarine in European countries is made with varying proportions of milk and water, American margarines are made with milk or skim milk only in the aqueous phase. The Federal (U.S.) margarine standard [27] also makes mandatory the culturing of the milk. Quite generally, cultured skim milk for margarine is cultured to an acidity of 0.80 to 0.85 per cent acid, calculated as lactic acid. Usually about one part of cultured milk or cultured milk plus sodium benzoate and salt is churned with four parts of margarine oil. Since soya-bean lecithin, monoglycerides and vitamins A and D concentrates are oil-miscible, these are commonly dissolved or dispersed in the margarine oil before churning. Since lecithin is readily hydrated and since sodium benzoate when dissolved in the cultured milk is readily converted to the oil-soluble benzoic acid, it is not surprising to learn that there is probably considerable interchange of benzoate and lecithin between aqueous and oil phases during the churning operation. In any event, the experimental evidence indicates that practically all of the sodium benzoate added to the cultured milk is extractable from the fat in the finished margarine as benzoic acid and that practically all of the phosphatide has found its way into the aqueous phase, as shown in the figures in Table 1.

Simple microscopic techniques plainly demonstrate that the margarine emulsion in the churn is a water-in-oil type; that is, that the milk phase is the internal phase. This is readily confirmed by the easy miscibility of the emulsion with oil (and its immiscibility with water). While it is difficult to produce equally acceptable evidence of this nature on the finished (solid) product, owing to the practically instantaneous chilling over the chill roll or in the "internal" dry cooling equipment (especially with the added emulsifying action of agitation during chilling), it seems

almost certain that the finished margarine is simply a solidified "water-in-oil" type of emulsion.

TABLE 1. BENZOIC ACID AND LECITHIN IN MARGARINE

	Benzoic Acid Found Calculated as % Sodium Benzoate
1. Whole margarine*	0.07
2. Margarine oil [separated from (1)]	0.06

	% Lecithin Found by Analysis
3. Margarine oil treated with 0.16% lecithin (before churning)	0.16
4. Margarine oil separated from margarine made with (3)	0.01–0.02

* Sodium benzoate added to cultured milk before churning.

Addition of Vitamin A

In the United States practically all margarine is fortified with vitamin A to the extent of over 15,000 U.S.P. XIII units per pound. Here again the control for maintaining this unitage is based on a physical measurement by means of the spectrophotometer, adequately supported by bioassay tests. The procedure, while simple in principle, requires careful sampling, as well as painstaking care in the laboratory. It consists of a spectrophotometric comparison of the identical margarine oils just before and after the addition of vitamin A with adequate agitation to insure uniform distribution of the vitamin A-bearing ingredient. Confirmatory bioassays are made in advance on the vitamin A-bearing addition and also subsequently on the finished product but, owing to the time required for bioassays, the day-to-day control of the vitamin A content of margarine must be carried out by spectrophotometric procedures. When margarine is colored with carotene the spectrophotometer is also employed for examination.

The Beckmann instrument is probably the most widely used for the routine control of vitamin A in margarine, although other spectrophotometers are also in use. Before the advent of the Beckmann instrument, the Hilger spectrophotometer was used and in the early 1930's, the earlier Hilger Vitameter found some application in control work.

Conclusion

Summarizing briefly then, the development of the margarine industry, from a technological standpoint, has been the most rapid in the entire history of the product since about 1930. While this was made possible, in part, by the advent of hydrogenation and modern deodorization

methods which came into being about the turn of the century, the most striking changes came some time after the Harkins-Langmuir concepts involving the physicochemical aspects of emulsification and their application. The latter, followed by additional mechanical handling developments in manufacture, not only made the finished product more satisfactory for consumer use, but also made possible vastly improved manufacturing procedures.

REFERENCES

1. Clayton, W., "Margarine," New York, Longmans, Green & Co., Inc., 1920.
2. Megé, H., U.S. Patent 146,012 (Dec. 30, 1873).
3. Lewkowitsch, J., "Chemical Technology and Analysis of Oils, Fats and Waxes," London, Macmillan & Co., Ltd., 1915.
4. Snodgrass, K., "Margarine as a Butter Substitute," Food Research Inst., Calif., Stanford Univ., 1930.
5. Brannt, W. T., "Animal and Vegetable Fats and Oils—Artificial Butters," Philadelphia, Baird & Co., 1888.
6. Storch, *Milch-Ztg.*, **19**, 304 (1890); Weigmann, *Milch-Ztg.*, **19**, 593 (1890); Conn, *Repts. Agr. Expt. Sta., Storrs, Conn.* (1890-98).
7. Hammer, B. W., "Dairy Bacteriology," John Wiley & Sons, Inc., 1928.
8. Van Niel, C. B., Kluyer, A. J., and Derx, H. G., *Biochem. Z.*, **ccx**, 234 (1929).
9. Schmalfuss, H., and Barthmeyer, H., *Ber.*, **60**, 1035 (1927).
10. Brit. Patent 46 (1905); Germ. Patent 175,381 (1906).
11. Eichstaedt, A., Richter, K., and Huettig, C., *Deut. Margarine Z.*, **16** (8), (1927); Steudel, H., and Massatsch, C., *ibid.*, **16** (13), (1927).
12. Sabatier, P., and Senderens, J. B., *Compt. rend.*, **132**, 210, 566, 1254 (1901); Sabatier, P., "Catalysts in Organic Chemistry," trans. E. E. Reid, New York, D. Van Nostrand Co., Inc., 1922.
13. Normann, W., Germ. Patent 141,029 (Aug. 14, 1902); Brit. Patent 1,515 (Nov. 26, 1903).
14. Weber, G. M., and Alsberg, C. L., "The American Vegetable Shortening Industry," Appendix B, Food Research Inst., Calif., Stanford Univ., 1934.
15. Bogue, R. H., "Colloidal Behavior," Vol. I, "Theory," New York, McGraw-Hill Book Co., Inc., 1924.
16. Davis, C. E., *Ind. Eng. Chem.*, **13**, 799 (1921).
17. Vahlteich, H. W., "Modern Margarine Manufacture," *Food Industries*, 436 (July, 1929).
18. Langmuir, I., *Chem. Met. Eng.*, **15**, 468 (1916).
19. Harkins, W. D., Brown, F. E., and Davies, E. C. H., *J. Am. Chem. Soc.*, **39**, 354 (1917); Harkins, W. D., Davies, and Clark, G. L., *ibid.*, 541 (1917).
20. Harkins, W. D., and Cheng, Y. C., *J. Am. Chem. Soc.*, **43**, 36 (1921); Harkins, W. D., Clark, G. L., and Roberts, L. E., *ibid.*, **42**, 700 (1920).
21. Harkins, W. D., and Roberts, L. E., *J. Am. Chem. Soc.*, **44**, 653 (1922).
22. Langmuir, I., *J. Am. Chem. Soc.*, **39**, 1848 (1917).
23. Luksch, E., *Die Seife*, **7**, 305 (1921) [from *Chem. Zentr.*, 278 (1922)].
24. Harris, B. R., *et al.*, U.S. Patents 1,917,249-259 incl. (1933).
25. Rosenbusch, R., and Reverey, G., *Margarine Industrie* (17) (1927); *Chem. Zentr.*, II, 2363 (1927); *Chem. Abstracts*, **22**, 4669 (1928).
26. Gooding, C. M., Neal, R. H., and Vahlteich, H. W., U.S. Patents 2,485,631-640 (Oct. 25, 1949); Brit. Patent Specification 625,683 (accepted July 1, 1949).
27. Federal Register (U.S.), p. 2762, June 7, 1941.

CHEMICAL WARFARE

Duncan MacRae

Technical Command, Army Chemical Center, Md.

IN 1931 Jerome Alexander [1] listed the principal chemical warfare materials substantially as follows:

War Gases. Mustard, lewisite, chloroacetophenone and bromobenzylcyanide.

Signalling Smokes. Red "paratoner"; yellow, chrysoidine plus auramine; blue, indigo; purple, indulin; and green, auramine plus indigo.

Screening Smokes. Phosphorus; sulfur trioxide (oleum); chlorosulfonic acid; tin, silicon, and titanium tetrachlorides; ammonium chloride, and zinc chloride (made by the interaction of zinc and carbon tetrachloride or hexachloroethane).

Irritant Smokes. Diphenylchloroarsine and diphenylaminechloroarsine.

Incendiaries. At that time incendiaries and flame-thrower fuels were held in so little esteem that they were not mentioned.

Weapons. Livens projector, 4-inch Stokes mortar, 4.2-inch chemical mortar and chemical artillery shell.

Gas-Mask Components. Activated charcoal and aerosol filters. At that time Alexander pointed out that the colloid-chemical aspects of chemical warfare included the use of aerosols in the form of toxic, screening, and signalling smokes and that the principal functions of the gas mask were to filter out aerosols and adsorb toxic vapors. Thus, a great part of the technique of chemical warfare was then based on the principles of colloid chemistry.

It is the purpose of the present paper to indicate the important published contributions of colloid chemistry to chemical warfare in the period from 1931 to 1949. During this period tremendous advances in the field of chemical warfare were made. War gases, including irritant smokes, were used to a comparatively small extent by the Italians in Abyssinia.[7] Signalling smokes were important and were widely used in combat.[4a] Screening smokes were employed on a scale never dreamed

of before in operations in both Europe and the Pacific.[4b] They were of
great value in connection with landing operations, they greatly reduced
losses in loading and unloading supplies at the docks, and were effective
in preventing the success of Kamikaze attacks by Japanese aviators.[4h]
The war economy of Japan had already been wrecked by incendiaries
two months before that country's surrender,[4f] and flame throwers greatly
accelerated the advance of MacArthur's troops northward by overcoming
very effective resistance by Japanese fortifications of the bunker type.[4g]
These great successes in chemical warfare were due in no small measure
to the application of the principles of colloid chemistry by military and
civilian scientists. The new chemical warfare materials and related de-
velopments in colloid chemistry are discussed below.

War Gases

In the field of war gases, the following compounds, not included in
the 1931 list, have been mentioned in authoritative publications[2, 4d] as
substances considered in preparations for gas warfare: nitrogen mustards,
$N(C_2H_4Cl)_3$; fluoroacetates, $CH_2FCOOCH_3$ and $CH_2FCOONa$; diiso-
propyl fluorophosphate, $(i\text{-}PrO)_2POF$. In addition to these, the two war
gases, hydrocyanic acid and cyanogen chloride, which had seen some
use in World War I, were perfected for use in World War II. These
compounds and other war gases are of interest to colloid chemists on
account of their chemical action on the colloidal components of living
organisms.

It has been found that tris (beta chloroethyl) amine undergoes cycliza-
tion in aqueous solution to form the quaternary imonium compound
$(ClC_2H_4)_2C_2H_4N^+Cl^-$. The following is quoted from Chapter XXXVI of
"Advances in Military Medicine"[2]:

"Whereas the original tertiary amine is a relatively nonreactive product
and, indeed, except for the property of cyclization may be said to be
chemically inert, the quaternary imonium compound with its ethyleni-
monium ring is one of the most reactive of organic chemicals. So reactive
is it that reaction occurs readily even with water, in the following
manner:

$$\underset{CH_2CH_2Cl}{\overset{\overset{\displaystyle CH_2CH_2}{\diagup\!\!\diagdown}}{{}^+N\!\!-\!CH_2CH_2Cl}} + H_2O \longrightarrow \underset{CH_2CH_2Cl}{\overset{\overset{\displaystyle CH_2CH_2OH}{\diagup}}{N\!\!-\!CH_2CH_2Cl}} + H^+$$
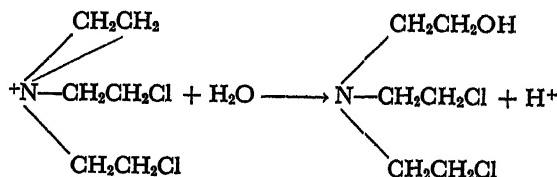
If substances other than water are present they can react competitively.
So reactive are some compounds that in their presence reaction with

water is negligible. As examples of compounds of biologic importance that readily react with the ethylenimonium ring of the nitrogen mustards may be mentioned the α-amino, imidazole, sulfide, phenolic, ε-amino, and imino groups of amino acids and peptides; inorganic phosphate, as well as glycerophosphate and hexose phosphates; the amino groups of adenosine and thiamine; and the pyridino-N of nicotinic acid amide and pyridoxine. In addition, reactions have been demonstrated to occur with proteins such as hemoglobin, insulin, gelatin, crystalline egg albumin, tobacco mosaic virus, ovalbumin, and protamine, as well as various purified enzymes. Thus it appears likely that the basis of the systemic toxicity of the nitrogen mustards resides in the high chemical reactivity of the ethylenimonium ring.

"The information afforded the chemists by the nitrogen mustards was the necessary tool for the elucidation of the chemistry of the sulfur mustards. Thus, in aqueous solutions sulfur mustard readily forms a sulfonium ring as follows:

$$S \begin{array}{c} \diagup CH_2CH_2Cl \\ \diagdown CH_2CH_2Cl \end{array} \longrightarrow {}^+S \begin{array}{c} \diagup \overset{\displaystyle CH_2CH_2}{\diagup} \\ \diagdown CH_2CH_2Cl \end{array} + Cl^-$$
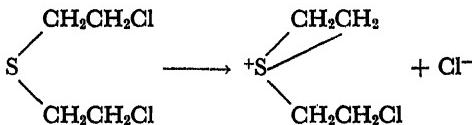
The ethylenesulfonium ring is even more reactive than the ethylenimonium."

In searching for the explanation of the basic mechanism of action of the fluoroacetates, attention was directed to their effects on the intermediary metabolism of carbohydrates. It was suggested that their specific action was to block the oxidation of acetate itself by competitive inhibition of the enzyme systems on which the utilization of acetate is dependent. While no fluoroacetate qualified as a war gas, sodium fluoroacetate was an outstanding success as a pesticide.[2a]

Reports indicating that quantities of an alkyl derivative of fluorophosphoric acid had been prepared in Germany led to the preparation of diisopropyl fluorophosphate in England. It has the property of reducing cholinesterase activity. It has been shown that rabbit and human serum, red cells and tissues, especially those of the liver, contain an enzyme that accelerates decomposition of fluorophosphates.[2d]

The classical observation that methemoglobin can combine with and thus effectively remove cyanide from its combination with cytochrome oxidase was amply confirmed.[2b]

The study of the mechanism of action of lewisite led to the discovery that a certain chemical compound could be used to prevent injury by vesicant arsenicals. That compound is 2,3-dimercaptopropanol and is known as BAL (British anti-lewisite). When the skin or eye of an animal

had been contaminated with lewisite, the application of an ointment containing BAL was very effective in minimizing the damage. The mechanism of its action is believed to be due to the fact that the —SH groups of BAL compete for the arsenic that would otherwise combine with similar groups in body proteins.[2c]

Screening Smokes

The principal addition to the list of screening smokes since 1931 is fog oil, a high-boiling petroleum fraction.[6a] Its success was due to the applications of the principles of colloid chemistry. With the aid of the Mie theory of light scattering [15] it was possible to determine the proper size of particle to give the maximum scattering in the region of visible light. For materials of refractive index of 1.5 this is a particle with a diameter of 0.6 micron. That theory was also useful in the development of apparatus for the rapid determination of particle size.[15] The conditions required for the production of a fog-oil aerosol with particles of this size were worked out in the laboratory and smoke generators designed and constructed to produce them. In some generators the oil was vaporized and the vapor allowed to escape through small nozzles into air. The size of the particle depended on the linear velocity of the vapor and the rate of cooling. These in turn were dependent on the rate of heating of the oil and the dimensions of the nozzles.[6a] In other generators a venturi throat was used for atomizing and mixing the oil with hot gases. Obviously the oil must have such vapor pressures as would permit it to be vaporized at a practicable temperature and yet be condensed into an aerosol at air temperatures prevalent in the field. Nor must it decompose excessively during the process of vaporization.

Theories relating to the rise and spread of aerosol clouds combined with knowledge of their ability to scatter light were useful in describing the field behavior of smoke screens. For example, when viewed from above, a smoke cloud from a single generator frequently has roughly the shape of a cigar. The length at the center line is obtained from the following equations:

$$X_{max} = \left(\frac{Q}{\sqrt{\pi} \, C_y \bar{U} N_0} \right)^{\frac{2}{2-n}}$$

The maximum half-width is given by the following:

$$Y_{max} = 0.248 \frac{Q}{\bar{U} N_0}$$

and

$$X \text{ (at max } y) = (0.606) \, X_{max}$$

in which [16]

Q = the source strength, g/sec
C_y = the turbulent diffusion coefficient in the crosswind direction
$\bar{U}$ = the mean wind velocity, m/sec
N_0 = the minimum density for screening, g/sq m
n = a number between 1 and 2 whose magnitude depends on atmospheric turbulence

Colored Signal Smokes. The following dyes have been found to produce satisfactory colored smoke clouds[12]:

Dye	Type	Color of Smoke
auramine	diphenylmethane	yellow
β-naphthaleneazodimethylaniline	azo	yellow
benzeneazodimethylaniline	azo	yellow
1-methylaminoanthraquinone	anthraquinone	red
9-diethylaminorosindone	azine	red
α-aminoanthraquinone *	anthraquinone	orange
1-amino-8-chloroanthraquinone *	anthraquinone	orange
quinizarin	anthraquinone	orange
1,4-diaminoanthraquinone †	anthraquinone	violet
1,4-diamino-2,3-dihydroanthraquinone †	anthraquinone	violet
1,5-di-p-toluidinoanthraquinone	anthraquinone	violet
1,8-di-p-toluidinoanthraquinone	anthraquinone	violet
1,4-di-p-toluidinoanthraquinone *	anthraquinone	green
1,4-dimethylaminoanthraquinone	anthraquinone	blue
1-hydroxy-4-p-toluidinoanthraquinone	anthraquinone	deep blue

Note: * Mixed with auramine for proper shade.
     † Mixed with 1-methylaminoanthraquinone for proper shade.

The successful use of these smoke clouds evidently depends on much the same principles as that of the screening smokes. The theory of best particle size for colored smokes does not seem to be so well-known.

Incendiaries

A widely used incendiary consisted of magnesium made into special bombs and clusters, and an even better incendiary involving the application of colloid science—the gelled gasolines.[11] In England it was found that gasolines gelled with rubber were a very effective incendiary for use in bombs because they were not dispersed into small particles by the burster charge, but rather in pieces weighing about one-half a pound, which proved to be very effective in setting fires. Owing to the scarcity of rubber, thickening materials were developed in the United States to replace rubber. The best known of these was a mixture of aluminum naphthenate and palmitate called "Napalm."

This material was not only effective for use in incendiary and fire bombs, but proved to be of great value when mixed with gasoline for use in flame throwers. Its high viscosity greatly increased the range of the flame thrower and its thixotropic properties [9] made this possible with a

minimum of pressure required to force the thickened gasoline through the hose and flame-thrower nozzle.

Weapons

In the recent war neither the Livens projector nor the 4-inch Stokes mortar were used. The 4.2-inch chemical mortar [4e] became one of the most effective weapons in support of infantry. Its range was increased to between 4000 and 6000 yards, and the fillings for its shell included not only phosphorus to produce screening smokes and colored smokes for signalling purposes, but also high explosives. The only chemical-warfare materials used in artillery shell were screening smokes and some colored smokes. Other important weapons used in World War II but not mentioned in the 1931 article are gas, smoke and incendiary bombs, flame throwers, grenades, smokes, generators and smoke pots. These are of little interest to colloid chemists aside from the materials used in them.

Gas Masks

The gas masks of World War II were quite different in many respects from those used in combat 25 years earlier. They were lighter and of different shape, but the adsorption of vapors and the filtration of aerosols still played a major part in their method of functioning.

The study of the mechanism of the filtration of aerosols, during the period covered by the present paper, resulted in vastly superior filters in the German gas masks. These were duplicated and improved in the American and British mask. The principles of canister design relating to the use of charcoal to absorb gases were greatly elaborated, and greatly improved charcoals were available as compared with those of 1931. The result of this is seen in the much smaller and lighter gas-mask canisters in modern gas masks. Corresponding improvements were made in collective protectors for the purification of large volumes of air.

Gas-Mask Charcoal. It was known prior to 1931 that the performance of a bed of charcoal in removing a vapor from an air stream by adsorption could be represented simply by an equation substantially equivalent to the following [17]:

$$C_0 QT = N_0 A (X - I) \qquad (1)$$

in which

C_0 = concentration of toxic gas in air stream at entrance face of bed
Q = rate of flow in volume per unit time
T = time
N_0 = saturation capacity of a unit gross volume of adsorbent for the (toxic) gas

> A = cross section of adsorbent bed
> X = thickness of bed
> I = critical bed depth, i.e., the actual intercept of a life-thickness
>   curve on the thickness axis

When $I=0$, the equation states that C_0QT, the weight of vapor adsorbed, is equal to M_0Ax, the value required to saturate the bed completely. This is nearly true for the process of adsorbing the vapor of a high-boiling liquid from dry air at very small rates of flow; but for useful flow rates, I has appreciable values and plays an important part in the design of gas-mask canisters.

Of the variables contained in equation (1), the colloid chemist is primarily interested in the saturation value N_0 and the constant I. The saturation value N_0 is a property of the charcoal which depends on the test concentration, C_0, and varies with that concentration; it can be read from an adsorption isotherm. One of the principal problems in gas-mask development is to provide charcoals with high values of N_0 for all toxic gases. In testing charcoals an important problem is to select a test gas and test conditions which will permit conclusions to be drawn as to the general serviceability of the charcoal in adsorbing all vapors. It has been known that N_0, expressed in volume of liquid per unit volume of adsorbent bed, in the case of many charcoals is nearly the same for all vapors at high relative pressures. (Sometimes, however, some of the charcoal capillaries are so small that substances with large molecules cannot enter them and therefore have smaller values of N_0 than substances with smaller molecules.) At lower relative pressures the adsorption isotherms of different substances on the same charcoal diverge widely. Recently progress has been made in estimating unknown adsorption isotherms from one that is known.[18, 19] In general, substances highly associated in liquid form are much less readily adsorbed by charcoal than are non-associated ones at the same relative pressure.[20]

Gas-mask charcoals not only serve to adsorb vapors, but also act as a support for specific chemical substances which react with certain gases and which also serve to catalyze gas reaction with water vapor or oxygen. Thus copper oxide has long been added to the charcoal to catalyze the oxidation of arsine. Its performance in this respect is improved by the presence of a small amount of silver salt.[6b] Copper oxide also reacts with hydrocyanic acid to form cuprous cyanide and cyanogen.[6c] Further, it reacts with hydrogen chloride formed by the hydrolysis of phosgene. The removal of cyanogen chloride is aided by the presence of pyridine or chromium salts. Many other impregnants of charcoal are also used to aid in the removal of these and other gases. In some cases a value of N_0 may be calculated from the amount of the substance which will react with one of the impregnants present. In others the processes are much

more complicated than simple adsorption and there does not appear to be any well-defined value of N_0 for the substance in question.

While N_0 is the absorption capacity of the charcoal at equilibrium, the constant I is related to the rate of adsorption. For most charcoal beds it is possible by varying the thickness of the bed to find a value at which the break concentration appears at the instant the test gas reaches the effluent side of the charcoal bed. This thickness is such that the test concentration, C_0, is reduced to the break concentration, C_e, in the time required for a point in the air stream to traverse the bed.

As the air stream flows through the irregular passages in the bed, the molecules of vapor diffuse out of the air stream through the thin film of stationary air that surrounds each charcoal particle, strike the solid particles and either stick or rebound. If the molecules do not rebound into the gas film, they may reevaporate, react with other molecules, or move over the surface. Even if every molecule that strikes the charcoal sticks and does not reevaporate later, it still takes a finite bed thickness to reduce the concentration from C_0 to C_e. Under these conditions

$$I = I_t := H_t^\text{s}\ln C_0/C_e \tag{2}$$

in which H_t is the thickness required to make $C_e = C_0/2.72$. H_t is what chemical engineers call the height of a transfer unit. For a bed packed with irregular particles it has been found by experiment that

$$H_t = KD_p(D_p\rho V/\mu)^{0.41}(\mu/\rho D_v)^{0.67} \tag{3}$$

Thus H_t depends on the particle diameter D_p; the linear air velocity, V; the molecular diffusion coefficient of the test gas, D_v; and the viscosity and density of air, μ and ρ respectively.[17]

So calculated, I_t is a lower limit of I based on the assumption that every molecule of test gas that hits the charcoal granules sticks and does not rebound or reevaporate later. (For in that case the concentration of test gas at the surface of the charcoal particle is zero.) This lower limit can be lowered by reducing the size of the charcoal granules, by decreasing the rate of flow of the air stream, or by using a test gas with a larger diffusion coefficient; but once the critical bed depth has reached this lower limit, it cannot be further reduced by any change in the interior structure of the charcoal or in the nature of its inner surfaces.

The critical bed depth has been found to be substantially equal to I_t in the case of chloropicrin and a well-activated charcoal; but in the case of the adsorption of water vapor by charcoal, the critical bed depth may be much greater than I_t. This is due to the fact that at low relative pressures (i.e., relative humidities in the case of water) well-activated charcoal adsorbs little, if any, water vapor. Consequently a large fraction of the number of molecules that strike the charcoal surface in a given time must

either rebound or quickly reevaporate. It may be even that under certain conditions all water molecules rebound or quickly reevaporate. Unless some adsorption takes place at C_e, I may be infinite.

It is evident from the above that in general

$$I = I_t + I_r \tag{4}$$

in which I_t is the part of the critical bed depth attributed to diffusion in the air stream and I_r, the part due to processes taking place in or on the surface of the granules.

Not only has the theory of the performance of gas-mask charcoal been extended; but progress has also been made in the practice of carbonizing suitable raw materials and activating the charcoal so obtained.[14] Thus the design and performance of that essential chemical-warfare item, the gas mask, depends to a large extent on a knowledge of the adsorption isotherm and of the kinetics of adsorption and desorption.

Gas Mask Filters. Toxic smokes were used in combat in World War I to penetrate the gas masks in use at that time. It then became necessary to add aerosol filters to the gas-mask canisters in addition to the charcoal already provided in them to remove toxic gases or vapors. The operation of these filters may be best understood if we first understand how aerosols penetrate a charcoal bed.

If we consider aerosol particles to be very large molecules, then a number of these particles with their Brownian motion in air constitute the vapor of a substance with a very high molecular weight. Such a substance would have a very low vapor pressure and once one of its molecules (aerosol particles) collided with a solid surface, it would stick to it and not reevaporate. This fact alone would tend to make aerosols less capable of penetrating charcoal beds than are vapor molecules. On the other hand, a vapor with such large molecules would have a much smaller diffusion coefficient than ordinary molecules. It is therefore evident from equations (2) and (3) above that, other things being equal, the critical bed depth would be greater for aerosols than for vapors.

Elsewhere,[21] it has been calculated that the diffusion coefficient of an aerosol composed of particles 0.302 micron in diameter, is 1.08×10^{-6} cm^2/sec as compared with 0.088 cm^2/sec for chloropicrin. Consequently, if the critical bed depth under certain conditions were 0.1 cm for chloropicrin, for the aerosol it would be:

$$I = 0.1 \text{ cm} \times (0.088/1.08 \times 10^{-6})^{2/3} = 190 \text{ cm}$$

It is evident that this value is much larger than would be practicable for use in any military gas mask. Consequently, existing gas masks would be penetrated at once by the aerosol. Or, viewed in another way, if the penetration of the charcoal bed by chloropicrin were 10^{-10}, that for the aerosol

in question would be the number whose logarithm is $(-10) \times (1.08 \times 10^{-6}/0.088)^{2/3} = -0.0056$. This corresponds to a penetration of about 98 per cent by the aerosol.

Equations (2) and (3) thus give an idea of the variables which affect the filtration of aerosols. However, there is more to the subject than this. For example, increase in the size of the aerosol particles leads to smaller diffusion coefficients and greater penetration; but it is not to be expected that this will continue indefinitely. As the particles increase in size they will be screened out mechanically or will be thrown out by their inertia when they suddenly change direction in the irregular passages of the filter.

The filters used in World War II were a great improvement over those available in 1931; but they still consisted largely of fibers. It was found that for a filter to give good protection against small particles, the fiber diameters should be approximately equal to those of the particles removed. Many fine fibers, notably those of asbestos, are used in aerosol filters.[6d]

An important contribution to the problem of filter development and test was the development of instruments for rapidly determining the size and concentration of the particles of the aerosol used in testing experimental and production filters.[6e]

Conclusion

In the period covered by the present paper, colloid chemistry has contributed greatly to chemical warfare. On the other hand, chemical warfare problems have greatly stimulated research in the field of colloid chemistry and papers resulting from that work are still appearing in the scientific literature.

REFERENCES

1. Alexander, J., in "Colloid Chemistry, Theoretical and Applied," ed. J. Alexander, Vol. III, pp. 239-50, New York, Reinhold Publishing Corp., 1931.
2. Andrus, E. C., ed., "Science in World War II: Advances in Military Medicine," Vol. II, pp. 546-620, Boston, Mass., Little, Brown & Co., 1948.
2a. *Ibid.*, pp. 552-4.
2b. *Ibid.*, p. 554.
2c. *Ibid.*, p. 556.
2d. *Ibid.*, p. 561.
3. Baxter, J. P., 3rd, "Scientists Against Time," Boston, Mass., Little, Brown & Co., 1946.
4. Chemical Corps Association, "The Chemical Warfare Service in World War II," p. 34, New York, Reinhold Publishing Corp., 1948.
4a. *Ibid.*, p. 78.
4b. *Ibid.*, Chap. IX.
4c. *Ibid.*, Chap. V.
4d. *Ibid.*, pp. 57-8.

4e. *Ibid.*, Chap. VIII.

4f. *Ibid.*, p. 69.

4g. *Ibid.*, Chap. X.

4h. *Ibid.*, p. 180.

5. *Chem. Revs.*, 44 (2), (1949)—ten articles on aerosols.

5a. Sinclair, D., and LaMer, V. K., *Chem. Revs.*, 44 (2), 245 (1949).

6. Noyes, W. A., Jr., ed., "Science in World War II: Chemistry," pp. 278-88, Boston, Mass., Little, Brown & Co., 1946.

6a. *Ibid.*, pp. 279-82.

6b. *Ibid.*, p. 306.

6c. *Ibid.*, p. 299.

6d. *Ibid.*, pp. 265-70.

6e. *Ibid.*, pp. 270-1.

7. Waitt, A. H., "Gas Warfare," p. 118, New York, Duell, Sloan & Pearce, Inc., 1942.

8. Williams, R. T., ed., "The Biochemical Reactions of Chemical Warfare Agents, Symposium held at London School of Hygiene and Tropical Medicine, Dec. 13, 1947," Cambridge, Engl., Cambridge Univ. Press, 1948.

9. Finkelstein, L., "Rheological Properties of Incendiary Gels," *J. Phys. Coll. Chem.*, 52, 1460-70 (1948).

10. Juhola, A. J., and Wiig, E. O., "Pore Structure in Activated Charcoal I and II," *J. Am. Chem. Soc.*, 71, 2069-80 (1949).

11. Rueggeberg, W. H. C., "Use of Aluminum Soaps and Other Fuel Thickeners in Gelling Gasolines," *J. Phys. Coll. Chem.*, 52, 1444-60 (1948).

12. Technical Command, "Colored Signal Smokes," *Chem. Eng. News*, 22, 1990-1 (1944).

13. Gucker, F. T., Jr., "Aerosols," *Sci. Monthly*, 48, 373-85 (1949).

14. Hormats, S., "German Process for Manufacturing Activated Charcoal," *Chem. Eng.* (1946).

15. Sinclair, D., and LaMer, V. K., "Light Scattering as a Measure of Particle Size in Aerosols," *Chem. Revs.*, 44, 245-67 (1949).

16. Sherwood, T. K., "The Geometry of Smoke Screens," *J. Meteorology*, 6, 416-9 (1949).

17. Klotz, I. M., "The Adsorption Wave," *Chem. Revs.*, 39, 241-67 (1946).

18. Dubinin, M. M., and Timofeev, D. P., *Chem. Abstracts*, 43, 16 (1949).

19. Emmett, P. H., "Adsorption and Pore-Size Measurements on Charcoals and Whetlerites," *Chem. Revs.*, 43, 69-148 (1948).

20. Volman, D. H., and Klotz, I. M., *J. Chem. Phys.*, 14, 642 (1946).

21. Katz, S. H., and MacRae, D., "Filtration of Aerosols by Granular Charcoal," *J. Phys. Coll. Chem.*, 52, 695-7 (1948).

PHOTOSENSITIVE GLASS

Donald Stookey

Corning Glass Works, Corning, N. Y.

Introduction

Colloidal particles are essential ingredients in certain colored glasses and in opal glasses. Gold-ruby, copper-ruby and silver-yellow glasses are colored by submicroscopic metal particles; selenium-ruby, by cadmium sulfoselenide crystals. Opal glass is rendered translucent by the presence of microscopic inclusions having refractive indices different from that of the glass. In general, these particles or inclusions are not present as such during the melting process, but precipitate from homogeneous solution * as the melt is cooled or reheated. This results in a high degree and uniformity of particle dispersion.

In June, 1947, Corning Glass Works announced the commercial production of both colored and opal glasses in which the formation and growth of the colloidal particles are controlled photographically. Such glasses are designated by the term "photosensitive glass." Two basic types exist: color transparencies in which the image consists of submicroscopic particles of gold, silver or copper; and photosensitive opals in which the image is made up of nonmetallic particles capable of diffusing light. Except for the photographic image, the glass remains clear and colorless. The photographs may be developed in many colors. They are three-dimensional, resulting in some cases in a stereoscopic illusion. The color transparencies are essentially grainless, since the coloring particles are too small to scatter light appreciably, and they are produced in an optically-homogeneous medium; hence the resolving power is very high. The photographs are believed to be as permanent as the glass articles themselves. Conventional glass-melting and -forming methods are employed in the manufacture of photosensitive glass, and any of the usual types of glassware may be made. The photographic process consists of exposure to actinic radiation, and development by heat-treatment.

* Evidence cited later in this paper indicates that gold, silver, and copper dissolve as ions rather than as atoms.

697

Research leading to the development of photosensitive glass was initiated by the discovery of Dalton [1, 2] in 1937 that the color of copper-ruby glass is caused to "strike in"—that is, to develop on heating the originally colorless glass—more readily when it has been exposed to ultra-violet light before heat-treatment.

In 1941, the author began a search for compositions in which the above phenomenon could be enhanced sufficiently to make photography in glass feasible. This was accomplished in the same year with glasses colored by copper.[3] Later work disclosed that superior results were obtainable by the use of gold, together with appropriate sensitizing agents and color modifiers.[4] More recently it was discovered that the photo-graphically-developed metal particles can act as nuclei for the formation and growth of nonmetallic crystals in certain glass compositions, result-ing in the photosensitive opal glasses. Meanwhile Armistead,[5] who had been investigating the coloration of glass by silver, was able to make photosensitive silver glasses. Detailed results are given in references [6, 7, 8] from which much of the material in the present paper has been taken.

Photosensitive Glass Compositions

Photosensitive glasses are very similar to certain conventional glasses in composition, except for minute additions of constituents that may be classified as photosensitive metals, optical sensitizers, and thermore-ducing agents. A general description is given below, and specific com-positions may be found in the patents cited.

Base Glass. Most conventional silicate-glass compositions have been found capable of photosensitivity, provided an appropriate combination of photosensitive ingredients is employed. Exceptions are those contain-ing appreciable quantities of lead or other strong absorbers of ultra-violet light. The presence of at least 5 per cent of alkali metal oxide appears necessary.

Glasses containing more than 5 per cent of barium oxide have been found especially advantageous when used as base glasses for gold because they permit a number of colors to be developed in addition to the usual gold-ruby produced in other glasses.

Photosensitive Metals. The more important photosensitive metals are gold, silver, and copper. Certain other metals—palladium, for example—were found capable of modifying the color when used in combination with one of the above-mentioned metals. It was determined that melting conditions should be oxidizing for gold- or silver-containing glasses, and mildly reducing for copper-containing glasses.

Sensitizers. *Thermoreducing Agents.* This type of sensitizer comprises compounds of tin or antimony.

The effect of adding traces of tin or antimony compounds to the batch is to increase the tendency of the metal to "strike" color on heat-treatment. Excessive quantities cause spontaneous coloration. In terms of the photographic effect, these compounds were found to decrease the exposure required to produce a latent image, but to reduce the contrast of the developed image.

Optical Sensitizers. This type of sensitizer is distinguished from the thermoreducing agents described above because the sensitizing effects were demonstrated to result from absorption of the activating radiation by the sensitizer. Their influence is shown by sensitization of the metal to new wavelengths absorbed by the sensitizer, or by more rapid photographic effects with no loss in contrast. Cerium is the most important optical sensitizer known at present. The effects of these substances are discussed below in the section on the latent image.

Photographic Process

The photographic process consists of two steps, exposure and development. Exposure is accomplished by irradiating the glass with either ionizing radiations or ultraviolet light in the band between 300 and 350 millimicrons. Use of the latter makes it practical to expose by the contact-print method through ordinary film or glass-plate negatives. Development of the color transparency consists of heating the glass at or above its annealing temperature until the image is sufficiently intense. Development of the image in a photosensitive opal glass requires a more complex heat-treatment, as will appear below. A separate "fixing" treatment is unnecessary because the image is permanent at ordinary temperatures. On development, the colors representing strong exposure appear more rapidly than those of weak exposure. Deeper layers of glass which have received less exposure than surface layers develop more slowly, so that penetration increases with heat-treatment.

Exposure. In order to produce a developable latent image, direct exposure of the order of two-milliwatt minutes per square centimeter in the effective spectral region is necessary. When ordinary film negatives are employed, the required exposure is from two to ten times more intense because of the light-absorption of the negative. Exposure conditions (light source, filters, and exposure time) determine the color, depth of penetration, intensity, and contrast of the image which can be obtained on subsequent development.

High-energy x-rays, alpha particles, and beta rays have been shown to produce developable images. It is hoped that photosensitive glass may serve in photography of the three-dimensional tracks of nuclear particles.

The Latent Image. The latent image which is formed within the glass

by exposure to actinic radiation is detected with certainty only by ob-
servation of the visible image subsequently developed. Formation of the
latent image usually is accompanied by a slight increase in the absorp-
tion of ultraviolet or visible radiations by the glass. However, this is not
a reliable measure of the latent image because the same change may be
produced in the absence of the photosensitive metal. Moreover, the
initial transmission is virtually restored if the glass is heated above
400°C, although a latent image persists at much higher temperatures.

A number of properties of the latent image have been determined
by subjecting it to various treatments and observing the visible image
after development.
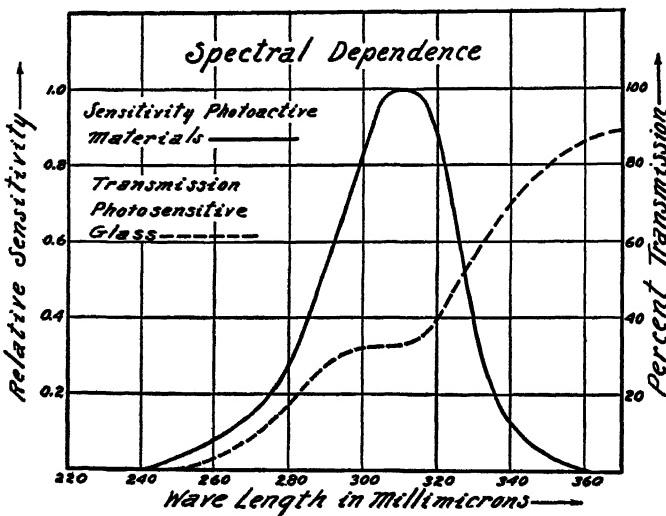
Figure 1. Spectral sensitivity curve of photosensitive ingredients. Per cent
transmission through 1-mm glass thickness.

Spectral Absorption. In common with most glasses, photosensitive
glasses may become slightly discolored as noted above, or "solarized," by
exposure to ultraviolet light or to ionizing radiation. Photosensitive
glasses containing silver or copper compounds have broad absorption
bands in the middle ultraviolet (300 to 370 mμ), which are characteristic
of these compounds and develop faint "solarization color"—blue-gray
for copper-containing glasses and yellow for silver-containing glasses—
after exposure. Absorption measurements made with a Beckman spectro-
photometer on photosensitive gold glasses show no absorption band that
can be attributed to the presence of gold compounds.

In photosensitive copper and silver glasses the latent image is formed
by exposure to ultraviolet light in the bands absorbed by the copper and

silver compounds. In photosensitive gold glass containing no added optical sensitizer, the latent image is formed only by short-wavelength ultraviolet light (254 mμ or shorter) which is absorbed by other ingredients of the glass, whereas addition of cerium compounds permits formation of a latent image by absorption of light wavelengths up to 350 mμ, in the band absorbed by cerous ions (see Figure 1).
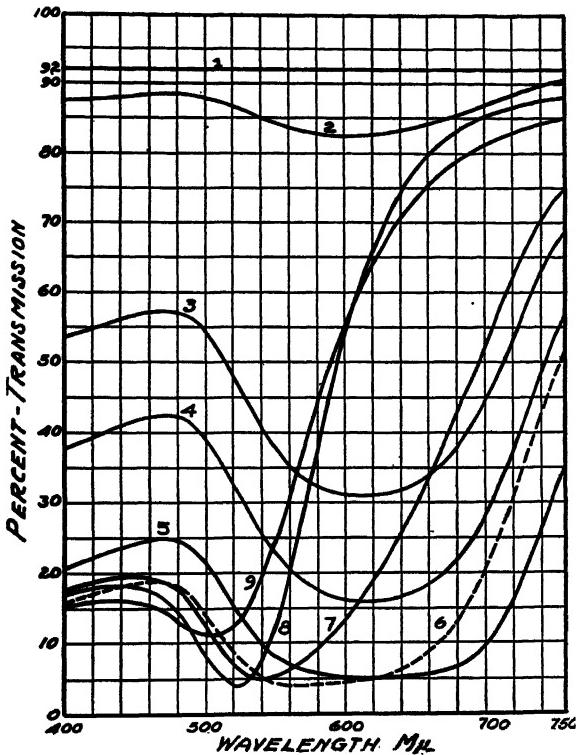
Figure 2. Spectral transmittances of developed colors in single plate of photosensitive gold glass 2-mm thick. Curves are numbered according to relative exposure time, except that 1 represents exposures of both 0 and 1.

Effects of Exposure Temperature and Heating Rate. No essential change in exposure effect is found in the range from liquid air temperatures up to about 580°C, the latter temperature (for the glass composition tested) being in the range where development occurs. It is possible to expose and develop color simultaneously at 580°C. Above 600°C, however, the glass may be exposed to ultraviolet light up to at least fifty times the normal exposure time without the production of a latent or a visible image.

The latent image may be destroyed almost instantaneously if the glass

is heated from room temperature to above 600°C. Destruction of the latent image is accompanied by thermoluminescence. Re-exposure of the piece in which the latent image has been destroyed results in formation of a normal latent image, as evidenced by subsequent normal development of a visible photograph.

Development. In order to produce a visible image after exposure, a photosensitive gold glass which has proved commercially practical is heated to a temperature between 580 and 650°C. Glasses with other viscosity-temperature relationships require corresponding developing temperatures. The development requires from 15 minutes to an hour, depending on the exposure and the type of color and depth effect desired. Inasmuch as the glass is somewhat deformable in this temperature range, special precautions must be taken not to mar the surface.

The color develops gradually, and development may be halted at any time by cooling the glass below the annealing temperature. Development is resumed with subsequent heating. The color and depth of penetration of the image depend primarily on the exposure conditions, but also to some extent on the development.

Development Colors in Photosensitive Barium—Gold Glass. Figure 2 gives the transmittance curves for a series of colors developed in a single plate of photosensitive barium-gold glass, as a function of exposure time. These colors range from pale blue through deep blue, purple, ruby and orange-red, with increasing exposure. The colors are the same by reflected and transmitted light, and the particles in all cases are too small to scatter an appreciable quantity of light.

Prolonged heat-treatment at high temperatures (650°C or above) tends to change the blue, purple and ruby colors to orange-red, so that curves 2 through 8 of Figure 2 all approach curve 9 after excessive heat-treatment.

Photosensitive Opal Glasses

It has been found possible, as stated previously, to employ the photographically-developed particles of gold, silver or copper as nuclei for the formation of nonmetallic crystals within the glass. This occurs in certain thermodynamically-unstable glasses in which one crystalline phase has a strong tendency to precipitate but is prevented from doing so by the high viscosity of the glass. The introduction of inhomogeneities (metal particles) by the photographic process enables controlled crystallization to occur in localized areas. Viscosity prevents the crystallization from spreading through the glass, and a photographic image is produced. In most cases the crystals are transparent and colorless (except for the color imparted by the metal), but differ from the glass in refractive index. It is

possible to decrease the concentration of metal sufficiently to eliminate any coloration, in which case the image is a light-diffusing white design within the clear glass.

It is interesting to note that a complex heat-treatment is required for the development of certain types of light-diffusing crystals. It is started by heating to a temperature above 600°C to develop metal particles which must exceed a limiting size before they can act as nuclei. Then the temperature must be decreased below 520°C to allow the light-diffusing crystal nuclei to form on the metallic nuclei, and again raised above 550°C to permit crystal growth. The number of nuclei can be closely controlled by the exposure and the first stage of the heat-treatment, while the crystal size depends upon the second heating cycle. Experiments indicate that each metal nucleus initiates growth of a single crystal. This behavior resembles, in some respects, that of a nuclear sol, which is a metal sol mixed with a solution containing metal ions and a reducing agent, furnishing nuclei for the growth of metal particles. Zsigmondy[9] employed this technique to determine the number of particles per unit volume in gold sols whose particles were too small to be observed ultramicroscopically. Adaptation of this method to photosensitive glasses has aided in a study of the concentration and distribution of the submicroscopic metal particles produced photographically.

If an excessive exposure is given, red coloration results from development but no light-diffusing crystals are formed. This indicates that the gold particles are too small to act as nuclei, or are so closely spaced that the dissolved material surrounding each metal particle is insufficient to form a crystal.

Theory

In a recent paper[8] evidence was presented which appears to prove that in glasses colored by colloidal gold, silver, or copper, the metal dissolves in an oxidized state during the melting process. In conventional gold- and copper-rubies and silver-yellow glasses, the dissolved coloring-metal compound subsequently becomes reduced to the metallic state as the glass is cooled or reheated. The reducing agent is a polyvalent ion such as selenium, tin, antimony, or arsenic, whose reduction potential compared to the metal ions is greater at low temperatures than at the melting temperature of glass (1400 to 1600°C). The metal, being insoluble, forms discrete particles which impart a characteristic color to the glass.

The photosensitive glasses described in the present work differ from the conventional metal-colored glasses in that they contain little or none of the polyvalent thermoreducing agents, hence remain colorless on heat-

treatment because the coloring metals remain in a soluble oxidized state. Instead of thermoreducing agents, these glasses contain optical sensitizers (among which the copper and silver compounds themselves are included) which become reducing agents for the dissolved coloring-metal compounds on exposure to actinic radiation.

The initial state of photosensitive glass, before exposure to actinic radiation, is therefore believed to be as follows: Ions of gold, silver, or cuprous copper are homogeneously dispersed in a hard glass matrix, in very low concentration, along with ions of the optical sensitizer. The glass at room temperature is rigid, permitting no mobility of ions or atoms and confining electronic mobility to a few atom diameters. The functions of other constituents of the glass are in general unimportant, except as they influence the oxidation state of the metal and sensitizer, or competitively absorb the actinic radiation.

The latent image is believed to consist of (1) photoelectrons emitted from light-sensitive ions such as silver, cuprous, cerous, or thallous, and held in a metastable activated state at trapping centers adjacent to the parent ions; and (2) metal ions capable of subsequently capturing the photoelectrons to form neutral atoms. The trapping centers may be metal ions or some "lattice imperfections" in the glass network. Because of the rigidity of the glass structure at room temperature, neither the electrical forces nor the ionic structure can be rearranged to new equilibrium states, so that the photochemical reaction (reduction of metal ions to the atomic state by the photoelectrons) is not completed until higher temperatures reduce the viscosity of the glass to the point where ionic diffusion occurs. If the exposure is made at high temperature, or if after exposure the glass is heated very rapidly to high temperature, violent thermal vibration destroys the latent image by returning the excited electrons to their original equilibrium state in the parent ions before the process of diffusion permits reaction with other ions to occur. This picture of the latent image is supported by the evidence of the presence of trapped photoelectrons, which is provided by the thermoluminescence of the irradiated glass, as well as by the behavior of the latent image on rapid heating or with high-temperature exposure.

Development by heating is believed to consist of two steps: capture of photoelectrons by metal ions to form atoms, and subsequent growth of metal particles, either by simple coalescence or by plating out of metal ions on contact with metal particles to which excess electrons have migrated.

Determination of the cause of the color variation in photosensitive barium-gold glasses presents an intriguing problem. According to Mie's optical theory of the color of metal sols [10] the color of particles such as those we are considering, whose size is below the light-scattering range,

depends only on selective absorption, which in turn depends on the chemical composition of the particle, its physical state, and its shape. In the case of gold, colors similar to those described above have been observed by others in colloidal solutions, and a very complete review of experimental and theoretical studies has been given by Zsigmondy in his book "Das Kolloide Gold." [9] He concludes that the yellow-red coloration is probably caused by gold particles in their highest degree of dispersion. (This presumably approaches atomic dispersion, since he made red hydrosols whose average particle diameter was only 16A, or 4 or 5 atoms in length.) With regard to the blue coloration, he cites evidence showing that it can be produced by colloidal aurous oxide, by aggregates of smaller gold particles, or by irregularly-shaped gold particles. The aggregated particles producing a blue color are in some cases smaller than massive crystals producing red color.

New experimental evidence on the subject, obtained in study of photosensitive glasses, is as follows:

(1) The developed color changes from blue through purple, ruby, and orange-red, with increasing exposure.

(2) The blue color develops more slowly than the red and is changed to red by longer heat-treatment.

(3) The number of developable gold particles increases with increasing exposure.

(4) Excessive exposure in a photosensitively-opacifiable glass results in red coloration without nucleation of opacifying crystals.

Since the concentration of gold in these glasses is very low, it appears reasonable to believe that the maximum particle size obtainable on development is inversely proportional to the number of gold particles per unit volume of glass. Hence, statements (1), (3), and (4) may be interpreted to mean that the smallest gold particles produce the orange-red color, and that the color shifts toward blue as particle size increases. However, in order to make this explanation consistent with (2), it is necessary to assume that the gold particles grow to a maximum size and then decrease in size with further heating. Such behavior appears somewhat improbable.

Other possible causes of the blue coloration are the following: amorphous particles, aggregation of gold particles, irregular particle shape, or colloidal aurous oxide. Any of these types of particles might conceivably be converted to regular gold crystals by excessive heating. Studies now being made with the electron microscope may eventually help to decide the true cause of the color variations.

Applications

Photosensitive glass possesses a unique combination of valuable properties, such as durability, dimensional stability, heat-resistance, transparency, and other characteristics inherent in glass. It can be made in the shape of tableware, jewelry, flat plate and glassware. Color transparencies in the form of plate glass are finding use in portrait and scenic photographs, photographic murals, decorative windows, church windows, advertising displays, ornamental tile, instrument dials, scales, and the like. Photosensitive opal glass containing three-dimensional grid patterns is employed in lighting fixtures for the purpose of controlling the direction of illumination and eliminating glare. It has many other potential applications of both a practical and decorative nature. In general photosensitive glass is a new photographic medium combining the properties of glass and the uses of photography into a new, many-sided tool for industry, art, and science.

REFERENCES

1. Dalton, R. H. (to Corning Glass Works), U.S. Patent 2,326,012 (Aug. 3, 1943).
2. Dalton, R. H. (to Corning Glass Works), U.S. Patent 2,422,472 (June 17, 1947).
3. Stookey, S. D. (to Corning Glass Works), Can. Patent 442,273 (June 17, 1947).
4. Stookey, S. D. (to Corning Glass Works), Can. Patent 444,616 (Oct. 21, 1947).
5. Armistead, W. H. (to Corning Glass Works), Can. Patent 442,272 (June 17, 1947).
6. Stookey, S. D., *J. Phot. Soc. Am.*, 14, 399-401 (1948).
7. ———, *Ind. Eng. Chem.*, 41 (4), 856-61 (1949).
8. ———, *J. Am. Ceramic Soc.*, 32 (8), 246-9 (1949).
9. Zsigmondy, R., and Thiessen, P. A., "Das Kolloide Gold," Leipzig, Akademische Verlagsgesellschaft, m. b. H., 1925; see also Zsigmondy, R., "Colloids and the Ultramicroscope," New York, Wiley & Sons, Inc., 1909; Engl. trans. of "Zur Erkenntnis der Kolloide," by J. Alexander, 1905.
10. Mie, G., *Ber. deut. physik. Ges.*, 5, 492-500 (1908).

AUTHOR INDEX

A

Abbott, A. D., 540
Abderhalden, E., 197
Abraham, Herbert, 510
Abraham, H. A., 554, 567
Achard, 323, 326
Adam, N. K., 110, 212, 223
Adams, E. Q., 148
Addison, C. C., 532, 540
Adelberg, E., 293
Adelsberger, L., 334, 344
Adler, E., 397
Aesculapius, 253
Agar, A., 223
Albaugh, F. W., 472
Albert, A., 223
Albutt, Sir T. C., 253
Alexander, A. E., 211 *et seq.*, 218, 223
Alexander, E. G., 323
Alexander, Jerome, 46, 81, 82, 110, 111,
    148, 182, 198, 201, 204, 208, 237, 243, 254,
    294, 312, 326, 344, 428, 429, 441, 472, 474,
    515, 520, 527, 649
Alexander, L. T., 441
Alexander-Jackson, E., 318, 320, 323, 325,
    327, 328
Allan, G., 223
Alpen, 310, 311
Ambrose, H. A., 472, 474
Am. Society for Testing Materials, 153
Anderson, C. D., 207
Anderson, John, 327
Anderson, P. C., 398
Anderson, T. G., 370, 398
Anderson-Kraft, C., 179
Andrade, E. N. da C., 110, 111, 151, 152,
    155, 157, 172
Andreas, J. M. 110, 311
Andresen, K. H., 472, 474
Andress, H. W., 473
Andress, K. R., 549
Andrews, 597
Annett, H. E., 298, 303
Ansbacher, S., 568
Anson, M. L., 223
Antonoff, G., 83, 86, 88, 92, 97, 109, 110,
    111 513
Appelt, K., 179
Aristotle, 206

A (continued)

Arloing, F., 318, 323, 325
Arloing, S., 323, 324
Arnow, L. E., 559, 568
Arrhenius, S., 152, 157, 167, 172
Asghar, A. G., 453
Astbury, W. T., 613
Astheimer, R. W., 108
Atterberg, 446
Austerweil, G., 185, 198
Avery, O. T., 293
Awapara, J., 198
Axelrod, J., 272, 276

B

Baart, M., 515
Babbitt, J. D., 626, 639
Babes, 324
Badger, 157, 167
Baer, von, 244, 245
Bailey, I. W., 622, 639
Bailey, J. R., 472
Baker, B. J., 107, 111
Baker, F., 157, 168, 173
Baker, M. O., 494
Baldwin, A., 616
Bale, W. F., 349, 396
Ball, J. M., 597
Ballantyne, M., 82
Ballou, G. A., 235
Bancroft, Wilder D., 110
Baptist, V. H., 198
Barkas, W. W., 625, 628, 630, 631, 639
Barnett, B., 497, 512
Barnett, S. J., 123
Barr, G. A., 152, 172
Barron, E. S. G., 270, 276
Barry, G. T., 198
Bart, R., 441
Bartell, F. E., 472
Bartlett, J. H., Jr., 32
Barton, K., 472
Bassett, E. G., 298, 304
Batchelder, A. C., 235
Bateman, L., 597
Bates, L. F., 123
Batschinski, A. J. Z., 154, 162, 172
Baudisch, O., 179
Bauverie, 324
Beadle, G. W., 280, 293, 299

SUBJECT INDEX

A

Acetanilide, 272
Acetate-butyrate films, 656
Acetate rayon, 607
Acetylcholine, 266
α-Acetylkynurine, 284
Achorion, 315
Acid formation in mouth,
  control of, 393
  rate of, 378, 379
  and sugar, 357
Acidizing, of oil wells, 455
Acidoids, in soils, 451
Acquired characteristics, 249, 250
Acridine, 459
Actinomyces, 318, 323
Actinomycetaceae, 323
Activators, in phosphors, 125, 136, 137
Active mass, 449
Adaptation in organisms, 257
Adaptive enzymes, 255
Adenyldiphosphate, 359 *et seq.*
Adenyltriphosphate, 359 *et seq.*
Adhesion, 69
Adhesion of "dirt," 538
Adsorbed monolayers, 531
Adsorption,
  of moisture by soils, 447
  multimolecular, 73
  of proteins, 216
  by wood, 624
Adsorption layer, in detergents, 521
Aerobacter aerogenes, 372, 373
Aerosols, 78
  agglomeration of, 404
  methods of collecting, 408
  oil fog, 688
  sound energy and, 413
  of toxic smokes, 692, 693
Aerosporin, 197
Age, and cosmetics, 557
Agglomeration of aersols, 403
Agglutination, of bacteria, 318
Agglutinin, 332
Agglutinogens, 330, 334, 338
Aggregates in liquids, 82, 105
Agnathia, 298
β-Alanine, in yeast, 260

Alcohol, 365
Alframin, 519
Allelic genes and blood groups, 337
Alloxazine nucleotides, 362
Alpha particles, energy of, 6
Amine reaction, with clay, 432
Amino acids,
  chromatography of, 181
  estimation of, 195
  preparation of, 200
p-Aminobenzoic acid, 271
p-Aminophenol, 272
Ammonium iodide, 75
Ammonium thioglycollate, 562
Analgesics, 269
Analysis, chromatographic, 183
Andrade function, 151, 152
Angle of contact, 92
Aniline-amylene, 103
Animalization of cellulose fibers, 608
Animals, variations in, 251
Anion structure, 233
Anionic soaps, 517
  classification of, 522
  factors governing, 524
Annual rings of trees, and past events, 620
Anthrax, immunity of chilled chickens to, 242
Antibiotics, 305
Antibodies, 242
  antigen-, 216
  mechanism of formation of, 243
  valency of, 341
Antigen-antibody, 216
Antigen films, 75, 76
Antiglobulin test, 342
Antiperspirants, 561
Anti-rhesus sera, 334
Anti-spattering, in margarine, 681
Antonoff's law, 84, 90, 92, 99, 103, 513, 514
Apatite, 349
Apomorphine, 269
Arctic Syntex A, 520
Arginine, 194, 285
Aromatics, cracking of, 490
Arsenic poisoning, 270
Ascaris lumbricoides, 221
Ascospores, in *Neurospora*, 279
Asexual phase, of bacteria, 313

726

Complement deficiency, 301
Compreg, 631
Compton scattering, 139
Conglutination, 332, 333
Conglutinin, 332
Conidia, in *Neurospora*, 279, 280
Conjugation, of drugs, 272
Conjugation reactions, 270, 271
Constitutive enzymes, 255
Consumer appeal, in packages, 661
Contact angle, 68
Copper,
  colloidal, 703
  complexes with protein, 227, 232
Cosmetics, 551
  classification, 565, 566
  effects of, 558
  preparation of, 566
Cotton, treated, 602
Cottrell precipitators, 404, 419, 487
Cottrell vs sonic systems, 426
Coupled reactions, 239
Crazing, 653
Creams, 559
Creation, 206
Critical micelle concentration, 532
Critical region, 103
Critical volume, in nuclear reactions, 19
Cross capillary analysis, 188
Crush resistance, 604
Cryochem, 665
Crystallites, 107
  in wood, 622
Crystals, structure of, 98
Culps medium, 369
Cultured milk, in margarine, 682
Cuprammonium rayon, 607
"Cuprophan," 648
Curie point, 113, 114, 115, 116, 123
Curium, 4
Curling, hair, 563
Cutis, 552
Cyanide, 269
Cyanogen chloride, 686
Cyanophile reticulum, 318
Cyanophiles, 314
Cybotaxis, 75
Cyclic diseases, relapses in, 323
Cysteine, 195
Cytochrome oxidase, 269

D

Dactylolysis, 298
Debye-Hückel equation, 54, 58
Decamethylene adipate, 169
Decibels, in sound, 415

Defects, heritable, 298
Deflocculating action of Graham's salt, 543
de Heen's theory, 109
Denaturation of proteins, 214
Density-temperature curves, 83, 104
Dental caries, 347 *et seq.*
  control of, 392
  factors in, 378
  mechanism of, 376
Dental plaque, 357, 386
Dentine, 348, 351
  bacteria in, 368
Deodorants, 561
Depth of surface zones, 67
Dermis, 552
Detergency, 529
Detergents,
  adsorption layer in, 521
  and chemical structure, 517, 529
  non-ionic, 534
  and sewage, 527
Deuterium, 20, 21
Development of exposed photosensitive glass, 702
Devulcanization, 571
Diamagnetism, 114
Diathesis, 251
Di-*n*-butyl succinate, 165
Diethylsuccinate, 167
Diffuse charges, 48
Diffusion, into cells, 556
Dihydroxyacetone phosphate, 361
Diisopropylfluorophosphate, 267, 274, 275, 686
Dilaudid, 269
Dilantin, 273
Dilatency, 432
Dimercaptopropanol, 242, 270
Dimethylfluorophosphate, 275
Di-*n*-octadecylsuccinate, 165
Diphenylhydantoin, 273
1,3-Diphosphoglyceric acid, 362
1,3-Diphosphoglyceric aldehyde, 361
Diphosphopyridine, 362
Dipole, magnetic, 114
Dipole moment, 101
Discontinuities in properties of liquids, 85
Disease,
  and drugs, 250
  host response to, 300
Dishumor, 657
Disintegration of atomic nuclei, 8, 12
Disorientation temperature, 74
Dispersion,
  of asphaltines, 505
  by Graham's salt, 546

Z